5

Designing Questions to Be Good Measures

In surveys, answers are of interest not intrinsically but because of their relationship to something they are supposed to measure. Good questions are reliable (providing consistent measures in comparable situations) and valid (answers correspond to what they are intended to measure). This chapter discusses theory and practical approaches to designing questions to be reliable and valid measures.

Designing a question for a survey instrument is designing a measure, not a conversational inquiry. In general, an answer given to a survey question is of no intrinsic interest. The answer is valuable to the extent that it can be shown to have a predictable relationship to facts or subjective states that are of interest. Good questions maximize the relationship between the answers recorded and what the researcher is trying to measure.

In one sense, survey answers are simply responses evoked in an artificial situation contrived by the researcher. The critical issue in this chapter is what an answer to a survey question tells us about some reality in which we have an interest. Let us look at a few specific kinds of answers and their meanings:

1. A respondent tells us that he voted for Dole rather than Clinton for president in 1996. The reality in which we are interested is which lever, if any, he pulled in the voting booth. The answer given in the survey may differ from what happened in the voting booth for any number of reasons. The respondent may have pulled the wrong lever and, therefore, not know for whom he really voted. The respondent could have forgotten for whom he voted. The respondent also could have altered his answer intentionally for some reason.
2. A respondent tells us how many times he went to the doctor for medical care during the past year. Is this the same number that the researcher would have come up with had he followed the respondent around for 24 hours every day during the past year? Problems of recall, of defining what constitutes a visit to a doctor, and of willingness to report accurately may affect the correspondence between the number the respondent gives and the count the researcher would have arrived at independently.
3. When a respondent rates her public school system as "good" rather than "fair" or "poor," the researcher will want to interpret this answer as reflecting evaluations and perceptions of that school system. If the respondent rated only one school (rather than the whole school system), tilted the answer to please the interviewer, or understood the question differently from others, her answer may not reflect the feelings the researcher tried to measure.

Many surveys are analyzed and interpreted as if the researcher knows for certain what the answer means. Studies designed to evaluate the correspondence between respondents' answers and true values show that many respondents answer many questions very well. Even so, to assume perfect correspondence between the answers people give and some other reality is naive. When answers are good measures, it is usually the result of careful design. In the following sections, specific ways that researchers can improve the correspondence between respondents' answers and the true state of affairs are discussed.

One goal of a good measure is to increase question reliability. When two respondents are in the same situation, they should answer the question in the same way. To the extent that there is inconsistency across respondents, random error is introduced, and the measurement is less precise. The first part of this chapter deals with how to increase the reliability of questions. There is also the issue of what a given answer means in relation to what a researcher is trying to measure: How well does the answer correspond? The latter two sections of this chapter are devoted to validity, the correspondence between answers and true values and ways to improve that correspondence (Cronbach & Meehl, 1955).

INCREASING THE RELIABILITY OF ANSWERS

One step toward ensuring consistent measurement is that each respondent in a sample is asked the same set of questions. Answers to these questions are recorded. The researcher would like to be able to make the assumption that differences in answers stem from differences among respondents rather than from differences in the stimuli to which respondents were exposed. The question's wording is obviously a central part of the stimulus.

- A survey data collection is an interaction between a researcher and a respondent. In a self-administered survey, the researcher speaks directly to the respondent through a written questionnaire or words on a computer screen. In other surveys, an interviewer reads the researcher's words to the respondent. In either case, the survey instrument is the protocol for one side of the interaction. In order to provide a consistent data collection experience for all respondents, a good question has the following properties:
  - The researcher's side of the question-and-answer process is entirely scripted, so that the questions as written fully prepare a respondent to answer questions.
  - The question means the same thing to every respondent.
  - The kinds of answers that constitute an appropriate response to the question are communicated consistently to all respondents.

Inadequate Wording

The simplest example of inadequate question wording is when, somehow, the researcher's words do not constitute a complete question.

INCOMPLETE WORDING

Bad	*Better*
5.1 Age?	What was your age on your last birthday?

Interviewers (or respondents) will have to add words or change words in order to make answerable questions from the words in the left column. If the goal is to have all respondents answering the same questions, then it is best if the researcher writes the questions fully.

Sometimes optional wording is required to fit differing respondent circumstances. That does not mean, however, that the researcher has to give up writing the questions. A common convention is to put optional wording in parentheses. These words will be used by the interviewer when they are appropriate to the situation and omitted when they are not needed.

EXAMPLES OF OPTIONAL WORDING

5.3 Were you (or anyone living here with you) attacked or beaten up by a stranger during the past year?

5.4 Did (you/he/she) report the attack to the police?

5.5 How old was (EACH PERSON) on (your/his/her) last birthday?

In example 5.3, the parenthetical phrase would be omitted if the interviewer already knew that the respondent lived alone. If more than one person lived in the household, though, the interviewer would include it. The parenthetical choice offered in 5.4 may seem minor. The parentheses, however, alert the interviewer to the fact that a wording choice must be made; the proper pronoun is used, and the principle is maintained that the interviewer need read only the questions exactly as written in order to present a satisfactory stimulus.

A variation that accomplishes the same thing is illustrated in 5.5. A format such as this might be used if the same question were to be used for each person in a household. Rather than repeat the identical words endlessly, a single question is written instructing the interviewer to substitute an appropriate designation (your husband/your son/your oldest daughter).

Of course, one advantage of computer-assisted instruments is that appropriate words can be filled in by the computer, rather than having interviewers adjust question wording to the circumstances. Whether on paper or via computer, the goal is to have the interviewer ask questions that make sense and take advantage of knowledge previously gained in the interview to tailor the questions to the respondent's individual circumstances. There is another kind of optional wording that is seen occasionally in questionnaires that is not acceptable.

EXAMPLE OF UNACCEPTABLE OPTIONAL WORDING

5.6 What do you like best about this neighborhood? (We're interested in anything, like houses, the people, the parks, or whatever.)

Presumably, this parenthetical probe was thought to be helpful to respondents who had difficulty in answering the question. From a measurement point of view, however, it undermines the principle of standardized interviewing. If interviewers use the parenthetical probe when a respondent does not readily come up with an answer, that subset of respondents will have answered a different question. Such optional probes usually are introduced when the researcher does not think the initial question is a very good one. The proper approach is to write a good question in the first place. Interviewers should not be given any options about what questions to read or how to read them except, as in the examples just discussed, to make the questions fit the circumstances of a particular respondent in a standardized way.

The following is a different example of incomplete question wording. There are three errors embedded in the example.

EXAMPLE OF POOR WORDING

5.7 I would like you to rate different features of your neighborhood as very good, good, fair, or poor. Please think carefully about each item as I read it.

a. Public schools

b. Parks

c. Public transportation

d. Other

The first problem with 5.7 is the order of the main stem. The response alternatives are read prior to an instruction to think carefully about the specific items. The respondent probably will forget the question. The interviewer likely will have to do some explaining or rewording before a respondent will be prepared to give an answer. Second, the words the interviewer needs to ask about the second item on the list, parks, are not provided. A much better question would be the following:

EXAMPLE OF BETTER WORDING

5.7a I am going to ask you to rate different features of your neighborhood. I want you to think carefully about your answers. How would you rate (FEATURE)—would you say very good, good, fair, or poor?

This gives the interviewer the wording needed for asking the first and all subsequent items on the list.

The third problem with the example is the fourth alternative, "other." What is the interviewer to say? Is he or she to make up some new question such as, "Is there anything else about your neighborhood you value?" How is the rating question to be worded? It is not uncommon to see "other" on a list of questions in a form similar to the example. Clearly, in the form presented in 5.7, the script is inadequate.

The examples given here illustrate questions that could not be presented consistently to all respondents as a result of incomplete wording. Another step needed to increase consistency is to create a set of questions that flow smoothly and easily. If questions have awkward or confusing wording, if there are words that are difficult to pronounce, or if combinations of words sound awkward together, interviewers will change the words to make the questions sound better or to make them easier to read. It may be possible to train and supervise interviewers to keep such changes to a minimum. Nevertheless, it only makes sense to help interviewers by giving them questions that are as easy to read as possible.

Ensuring Consistent Meaning to All Respondents

If all respondents are asked exactly the same questions, one step has been taken to ensure that differences in answers can be attributed to differences in respondents. But there is a further consideration: The questions should all mean the same thing to all respondents. If two respondents understand the question to mean different things, their answers may be different for that reason alone.

One potential problem is using words that are not understood universally. In general samples, it is important to remember that a range of educational experiences and cultural backgrounds will be represented. Even with well-educated respondents, using simple words that are short and understood widely is a sound approach to questionnaire design.

Undoubtedly, a much more common error than using unfamiliar words is the use of terms or concepts that can have multiple meanings. The prevalence of misunderstanding of common wording has been well documented by those who have studied the problem (e.g., Belson, 1981; Fowler, 1992; Oksenberg, Cannell, & Kalton, 1991; Tanur, 1991).

POORLY DEFINED TERMS

5.8 How many times in the past year have you seen or talked with a doctor about your health?

Problem. There are two ambiguous terms or concepts in this question. First, there is basis for uncertainty about what constitutes a doctor. Are only people practicing medicine with M.D. degrees included? If so, then psychiatrists are included, but psychologists, chiropractors, osteopaths, and podiatrists are not. What about physicians' assistants or nurses who work directly for doctors in doctors' offices? If a person goes to a doctor's office for an inoculation that is given by a nurse, does this count?

Second, what constitutes seeing or talking with a doctor? Do telephone consultations count? Do visits to a doctor's office when the doctor is not seen count?

Solutions. Often the best approach is to provide respondents and interviewers with the definitions they need.

5.8a We are going to ask about visits to doctors and getting medical advice from doctors. In this case, we are interested in all professional personnel who have M.D. degrees or work directly for an M.D. in the office, such as a nurse or medical assistant.

When the definition of what is wanted is extremely complicated and would take a very long time to define, as may be the case in this question, an additional constructive approach may be to ask supplementary questions about desired events that are particularly likely to be omitted. For example, visits to psychiatrists, visits for inoculations, and telephone consultations often are underreported and may warrant special follow-up questions.

POORLY DEFINED TERMS

5.9 Did you eat breakfast yesterday?

Problem. The difficulty is that the definition of breakfast varies widely. Some people consider coffee and a donut anytime before noon to be breakfast. Others do not consider that they have had breakfast unless it includes a major entree, such as bacon and eggs, and is consumed before 8 a.m. If the objective is to measure morning food consumption, the results are likely to contain considerable error stemming from differing definitions of breakfast.

Solutions. There are two approaches to the solution. On the one hand, one might choose to define breakfast:

5.9a For our purposes, let us consider breakfast to be a meal, eaten before 10:00 in the morning, that includes some protein such as eggs, meat, or milk, some grain such as toast or cereal, and some fruit or vegetable. Using that definition, did you have breakfast yesterday?

Although this often is a very good approach, in this case it is very complicated. Instead of trying to communicate a common definition to respondents, the researcher may simply ask people to report what they consumed before 10 a.m. At the coding stage, what was eaten can be evaluated consistently to see if it meets the standards for breakfast, without requiring each respondent to share the same definition.

POORLY DEFINED TERMS

5.10 Do you favor or oppose gun control legislation?

Problem. Gun control legislation can mean banning the legal sale of certain kinds of guns, asking people to register their guns, limiting the number or the kinds of guns that people may possess, or limiting which people may possess them. Answers cannot be interpreted without assumptions about what respondents think the question means. Respondents will undoubtedly interpret this question differently.

5.10a One proposal for the control of guns is that no person who ever had been convicted of a violent crime would be allowed to purchase or own a pistol, rifle, or shotgun. Would you oppose or support legislation like that?

One could argue that this is only one of a variety of proposals for gun control. That is exactly the point. If one wants to ask multiple questions about different possible strategies for gun control, one should ask separate specific questions that can be understood commonly by all respondents and interpreted by researchers. One does not solve the problem of a complex issue by leaving it to the respondents to decide what question they want to answer.

There is a potential tension between providing a complicated definition to all respondents and trying to keep questions clear and simple. This is particularly true for interviewer-administered surveys, as long definitions are particularly hard to grasp when they are delivered orally.

A potential approach is to tell interviewers to provide definitions to respondents who ask for clarification or appear to misunderstand a question. One concern about such approaches is that interviewers will not give consistent definitions if they have to improvise. However, computer-assisted interviewing makes it easy to provide interviewers with a precisely worded definition. The other, more important, concern is that only some respondents will get the needed definition. Those respondents who do not ask for clarification or do not appear confused will lack important information that might affect their answers.

Conrad and Schober (2000) experimented with giving interviewers freedom to provide definitions and explanations when they seemed needed. There was some evidence that accuracy improved, but the increases came at a price of more interviewer training and longer interviews. While there is need for more research on how to ask questions about complex concepts, the general approach of avoiding complex or ambiguous terms, and defining those that are used in the question wording, is the best approach for most surveys.

AVOIDING MULTIPLE QUESTIONS

Another way to make questions unreliable is to ask two questions at once.

5.11 Do you want to be rich and famous?

The problem is obvious: *Rich* and *famous* are not the same. A person could want to be one but not the other. Respondents, when faced with two questions, will have to decide which to answer, and that decision will be made inconsistently by different respondents.

Most multiple questions are somewhat subtler, however.

5.12 In the last 30 days, when you withdrew cash from an ATM machine, how often did you withdraw less than $25—always, usually, sometimes, never?

This question requires three cognitive calculations: calculate the number of visits to an ATM machine, the number of times less than $25 was withdrawn, and the relationship between the two numbers. While technically there is only one question, it is necessary to answer at least two prior questions in order to produce the answer. It would be better question design to use two questions.

5.12a In the last 30 days, how many times did you withdraw cash from an ATM machine?

5.12b (IF ANY) On how many of those times did you withdraw less than $25?

Note two other virtues of the 5.12a and 5.12b series. First, it identifies those who did not use an ATM machine at all, to whom the question does not apply. Second, by asking for numbers in both questions, it avoids having respondents do a calculation. Simplifying the demands on respondents is almost always a good idea.

5.13 To what kind of place do you go to for your routine medical care?

This question assumes that all respondents get routine medical care, which is not an accurate assumption. It should be asked as two questions. Probably the best approach is to ask if the respondent has gotten any routine medical care in some period—for example, the past 12 months. If so, follow with a question about the kind of place.

The "Don't Know" Option

When respondents are being asked questions about their own lives, feelings, or experiences, a "don't know" response is often a statement that they are unwilling to do the work required to give an answer. On the other hand, sometimes we ask respondents questions concerning things about which they legitimately do not know. As the subject of the questions gets farther from their immediate lives, the more plausible and reasonable it is that some respondents will not have adequate knowledge on which to base an answer or will not have formed an opinion or feeling. In those cases, we have another example of a question that actually is two questions at once: do you have the information needed to answer the question, and, if so, what is the answer?

There are two approaches to dealing with such a possibility. One simply can ask the questions of all respondents, relying on the respondent to volunteer a "don't know" answer. Respondents differ in their willingness to volunteer that they "don't know," however (Schuman & Presser, 1981), and interviewers are inconsistent in how they handle "don't know" responses (Fowler & Mangione, 1990; Groves, 1989). The alternative is to ask all respondents a standardized screening question about whether or not they feel familiar enough with a topic to have an opinion or feeling about it.

When a researcher is dealing with a topic about which familiarity is high, whether or not a screening question for knowledge is asked is probably not important. When a notable number of respondents will not be familiar with, or have not thought about, whatever the question is dealing with, it probably is best to ask a screening question about familiarity with the topic.

Specialized Wording for Special Subgroups

Researchers have wrestled with the fact that the vocabularies in different subgroups of the population are not the same. One could argue that standardized measurement actually would require different questions for different subgroups (Schaeffer, 1992).

Designing different forms of questionnaires for different subgroups, however, is almost never done. Rather, methodologists tend to work very hard to attempt to find wording that has consistent meaning across an entire population. Even though there are situations where a question's wording is more typical of the speech of one segment of a community than another (most often the better-educated segment), finding exactly comparable words for some other group of the population and then giving interviewers reliable rules for deciding when to ask which version is so difficult that it is likely to produce more unreliability than it eliminates.

The extreme challenge is how to collect comparable data from people who speak different languages. The most careful efforts translate an original version into the new language, have a different translator back translate the new version into the original language, and then try to reconcile the differences between the original and the back-translated version.

This process can be greatly improved if the designers of the original questions were concerned about ease of translation. For example, numbers translate more readily across languages than adjectives. Abstract concepts and words that are colloquial are likely to be particularly hard to translate accurately. Even when great care is taken, it is very hard to be sure people are answering comparable questions across languages. It is doubtful that adjectival rating scales are ever comparable across languages. The more concrete the questions, the better the chances for comparability of results across languages or cultures. Marin and Marin (1991) present a good analysis of the challenges of collecting comparable data from English- and Spanish-speaking people.

Standardized Expectations for Type of Response

As stated, it is important to give interviewers a good script so that they can read the questions exactly as worded, and it is important to design questions that mean the same thing to all respondents. The other key component of a good question is that respondents should have the same perception of what constitutes an adequate answer for the question.

The simplest way to give respondents the same perceptions of what constitutes an adequate answer is to provide them with a list of acceptable answers. Such questions are called closed questions. The respondent has to choose one, or sometimes more than one, of a set of alternatives provided by the researcher.

Closed questions are not suitable in all instances. The range of possible answers may be more extensive than it is reasonable to provide. The researcher may not feel that all reasonable answers can be anticipated. For such reasons, the researcher may prefer not to provide a list of alternatives to the respondent. In that case, the question must communicate the kind of response wanted as well as possible.

5.14 When did you have the measles?

Problem. The question does not specify the terms in which the respondent is to answer. Consider the following possible answers: "Five years ago"; "While I was in the army"; "When I was pregnant with our first child"; "When I was 32"; "In 1987." All of these answers could be given by the same person, and all are appropriate answers to the question as posed. They are not all acceptable in the same survey, however, because descriptive statistics require comparable answers. An interviewer cannot use the words in example 5.14 and consistently obtain comparable data, because each respondent must guess what kind of answer is wanted.

Solution. A new question must be created that explains to the respondent what kind of answer is wanted.

5.14a How old were you when you had the measles?

Obviously, 5.14a is the way the question should have been worded by the researcher for all respondents.

5.15 Why did you vote for Candidate A?

Problems. Almost all "why" questions pose problems. The reason is that one's sense of causality or frame of reference can influence answers. In the particular instance described, the respondent may choose to talk about the strengths of Candidate A, the weaknesses of Candidate B, or the reasons he or she used certain criteria (My mother was a lifelong Republican). Hence respondents who see things exactly the same way may answer differently.

Solution. Specify the focus of the answer:

5.15a What characteristics of Candidate A led you to vote for (him/her) over Candidate B?

Such a question explains to respondents that the researcher wants them to talk about Candidate A, the person for whom they voted. If all respondents answer with that same frame of reference, the researcher then will be able to compare responses from different respondents in a direct fashion.

5.16 What are some of the things about this neighborhood that you like best?

Problems. In response to a question like this, some people will make only one or two points, whereas others will make many. It is possible that such differences reflect important differences in respondent perceptions or feelings. Research has shown pretty clearly, however, that education is related highly to the number of answers people give to such questions. Interviewers also affect the number of answers.

Solution. Specify the number of points to be made:

5.16a What is the feature of this neighborhood that you would single out as the one you like most?

5.16b Tell me the three things about this neighborhood that you like most about living here.

Although this may not be a satisfactory solution for all questions, for many such questions, it is an effective way of reducing unwanted variation in answers across respondents.

The basic point is that answers can vary because respondents have a different understanding of the kind of responses that are appropriate. Better specification of the properties of the answer desired can remove a needless source of unreliability in the measurement process.

TYPES OF MEASURES/TYPES OF QUESTIONS

Introduction

These procedures are designed to maximize reliability, the extent to which people in comparable situations will answer questions in similar ways. One can measure with perfect reliability, though, and still not be measuring what one wants to measure. The extent to which the answer given is a true measure and means what the researcher wants or expects it to mean is called validity. In this section, aspects of the design of questions are discussed, in addition to steps to maximize the reliability of questions, that can increase the validity of survey measures.

For this discussion, it is necessary to distinguish between questions designed to measure facts or objectively measurable events and questions designed to measure subjective states such as attitudes, opinions, and feelings. Even though there are questions that fall in a murky area on the border between these two categories, the idea of validity is somewhat different for objective and subjective measures.

If it is possible to check the accuracy of an answer by some independent observation, then the measure of validity becomes the similarity of the survey report to the value of some "true" measure. In theory, one could obtain an independent, accurate count of the number of times that an individual used an ATM during a year. Although in practice it may be very difficult to obtain such an independent measure (e.g., getting access to the relevant records could be impossible), the understanding of validity can be consistent for objective situations.

In contrast, when people are asked about subjective states, feelings, attitudes, and opinions, there is no objective way of validating the answers. Only the respondent has access to his or her feelings and opinions. Thus the validity of reports of subjective states can be assessed only by their correlations with other answers that a person gives or with other facts about the respondent's life that one thinks should be related to what is being measured. For such measures, there is no truly independent direct measure possible; the meaning of answers must be inferred from patterns of association.

Levels of Measurement

There are four different ways in which measurement is carried out in the social sciences. This produces four different kinds of tasks for respondents and four different kinds of data for analysis:

Nominal: People or events are sorted into unordered categories (Are you male or female?).

Ordinal: People or events are ordered or placed in ordered categories along a single dimension (How would you rate your health—very good, good, fair, or poor?).

Interval data: Numbers are attached that provide meaningful information about the distance between ordered stimuli or classes (in fact, interval data are very rare; Fahrenheit temperature is one of the few common examples).

Ratio data: Numbers are assigned such that ratios between values are meaningful, as well as the intervals between them. Common examples are counts or measurements by an objective, physical scale such as distance, weight, or pressure (How old were you on your last birthday?).

Most often in surveys, when one is collecting factual data, respondents are asked to fit themselves or their experiences into a category, creating nominal data, or they are asked for a number, most often ratio data. "Are you employed?", "Are you married?", and "Do you have arthritis?" are examples of questions that provide nominal data. "How many times have you seen a doctor?", "How much do you weigh?", and "What is the hourly rate you are paid?" are examples of questions that ask respondents to provide real numbers for ratio data.

When gathering factual data, respondents may be asked for ordinal answers. For example, they may be asked to report their incomes in relatively large categories or to describe their behavior in nonnumerical terms (e.g., usually, occasionally, seldom, or never). When respondents are asked to report factual events in ordinal terms, it is because great precision is not required by the researcher or because the task of reporting an exact number is considered too difficult. There usually is a real numerical basis, however, underlying an ordinal answer to a factual question.

The situation is somewhat different with respect to reports of subjective data. Although there have been efforts over the years, first in the work of psycho-physical psychologists (e.g., Thurstone & Chave, 1929), to have people assign numbers to subjective states that met the assumptions of interval and ratio data, for the most part respondents are asked to provide nominal and ordinal data about subjective states. The nominal question is, "Into which category do your feelings, opinions, or perceptions fall?" The ordinal question is, "Where along this continuum do your feelings, opinions, or perceptions fall?"

When designing a survey instrument, a basic task of the researcher is to decide what kind of measurement is desired. When that decision is made, there are some clear implications for the form in which the question will be asked.

Types of Questions

Survey questions can be classified roughly into two groups: those for which a list of acceptable responses is provided to the respondent (closed

questions) and those for which the acceptable responses are not provided exactly to the respondent (open questions).

When the goal is to put people in unordered categories (nominal data), the researcher has a choice about whether to ask an open or closed question. Virtually identical questions can be designed in either form.

EXAMPLES OF OPEN AND CLOSED QUESTIONS

5.17 What health conditions do you have? (open)

5.17a Which of the following conditions do you currently have? (READ LIST) (closed)

5.18 What do you consider to be the most important problem facing our country today? (open)

5.18a Here is a list of problems that many people in the country are concerned about. Which do you consider to be the most important problem facing your country today? (closed)

There are advantages to open questions. They permit the researcher to obtain answers that were unanticipated. They also may describe more closely the real views of the respondents. Third, and this is not a trivial point, respondents like the opportunity to answer some questions in their own words. To answer only by choosing a provided response and never to have an opportunity to say what is on one's mind can be a frustrating experience. Finally, open questions are appropriate when the list of possible answers is longer than is feasible to present to respondents.

Despite all this, however, closed questions are usually a more satisfactory way of creating data. There are three reasons for this:

1. The respondent can perform more reliably the task of answering the question when response alternatives are given.
2. The researcher can perform more reliably the task of interpreting the meaning of answers when the alternatives are given to the respondent (Schuman & Presser, 1981).
3. When a completely open question is asked, many people give relatively rare answers that are not analytically useful. Providing respondents with a constrained number of answer options increases the likelihood that there will be enough people giving any particular answer to be analytically interesting.

Finally, if the researcher wants ordinal data, the categories must be provided to the respondent. One cannot order responses reliably along a single continuum unless a set of permissible ordered answers is specified in the question. Further discussion about the task that is given to respondents when they are asked to perform an ordinal task is appropriate, because it is probably the most prevalent kind of measurement in survey research.

Figure 5.1 shows a continuum (this case concerns having respondents make a rating of some sort, but the general approach applies to all ordinal questions). There is a dimension assumed by the researcher that goes from the most negative feelings possible to the most positive feelings possible. The way survey researchers get respondents into ordered categories is to put designations or labels on such a continuum. Respondents then are asked to consider the labels, consider their own feelings or opinions, and place themselves in the proper category.

There are two points worth making about the kinds of data that result from such questions. First, respondents will differ in their understanding of what the labels or categories mean. The only assumption that is necessary in order to make meaningful analyses, however, is that, on the average, the people who rate their feelings as "good" feel more positively than those who rate their feelings as "fair." To the extent that people differ some in their understanding of and criteria for "good" and "fair," there is unreliability in the measurement, but the measurement will still have meaning (i.e., correlate with the underlying feeling state that the researcher wants to measure).

Second, an ordinal scale measurement like this is relative. The distribution of people choosing a particular label or category depends on the particular scale that is presented.

Consider the rating scale in Figure 5.1 again and consider two approaches to creating ordinal scales. In one case, the researcher used a 3-point scale: good, fair, and poor. In the second case, the researcher used five descriptive options: excellent, very good, good, fair, and poor. When one compares the two scales, one can see that adding "excellent" and "very good" in all probability does not simply break up the "good" category into three pieces. Rather, it changes the whole sense of the scale. People respond to the ordinal position of categories as well as to the descriptors. "Fair" almost certainly is farther to the negative side of the continuum when it is the fourth point on the scale than when it is the second. Thus one would expect more people to give a rating of "good" or better with the 5-point scale than with the 3-point scale.

FEELING ABOUT SOMETHING

Extremely Positive Extremely Negative

TWO-CATEGORY SCALE

Good Not Good

THREE-CATEGORY SCALE

Good Fair Poor

FOUR-CATEGORY SCALE

Very Good Good Fair Poor

FIVE-CATEGORY SCALE

Excellent Very Good Good Fair Poor

Such scales are meaningful if used as they are supposed to be used: to order people. By itself, however, a statement that some percentage of the population feels something is "good or better" is not appropriate, because

it implies that the population is being described in some absolute sense. In fact, the percentage would change if the question were different. Only comparative statements (or statements about relationships) are justifiable when one is using ordinal measures:

- comparing answers to the same question across groups (e.g., 20% more of those in group A than in group B rated the candidate as "good or better")
- comparing answers from comparable samples over time (e.g., 10% more rated the candidate "good" or better in January than did so in November)

The same general comments apply to data obtained by having respondents order items (e.g., Consider the schools, police services, and trash collection. Which is the most important city service to you?). The percentage giving any item top ranking, or the average ranking of an item, is completely dependent on the particular list provided. Comparisons between distributions when the alternatives have been changed at all are not meaningful.

Agree-Disagree Items: A Special Case

Agree-disagree items are very prevalent in survey research and therefore deserve special attention. The task that respondents are given in such items is different from that of placing themselves in an ordered category. The usual approach is to read a statement to respondents and to ask them if they agree or disagree with that statement. The statement is located somewhere on a continuum such as that portrayed in Figure 5.1. Respondents' locations on that continuum are calculated by figuring out whether they say their feelings are very close to that statement (by agreeing) or are very far from where that statement is located (by disagreeing).

When one compares questions posed in the agree-disagree format with questions in the straightforward rating format, there are numerous disadvantages to the former. Compare the following:

5.19 My health is poor. Do you strongly agree, agree, disagree, or strongly disagree?

5.19a How would you rate your health—excellent, very good, good, fair, or poor?

The disadvantages to the first statement are as follows:

- The rating scale sorts respondents into five categories; the agree-disagree question is almost always analyzed by putting respondents into two groups (agrees or disagrees). Hence more information is gained from the rating.
- Agree-disagree questions, in order to be interpretable, can be asked only about extremes of a continuum. If the statement was, "My health is fair," a person could disagree either because it was "good" or because it was "poor." This feature limits the ability to order people in the middle of a continuum.
- Respondents often find it confusing that the way to say their health is good is to disagree that their health is poor.
- Studies show that some respondents are particularly likely to agree (or acquiesce) when questions are put in this form; that is, there are people who would agree both that their health is "poor" and that it is "not poor" if question 5.19 was stated in the negative (Dillman & Tarnai, 1991; Schuman & Presser, 1981).

For unidimensional scaling tasks, it is hard to justify using 5.19 rather than 5.19a. A very common usage of the format, however, is to obtain responses to complex statements such as the following:

5.20 With economic conditions the way they are these days, it really isn't fair to have more than two children.

This question is asking about at least three things at once: the perceived state of the economy, views on the appropriate maximum number of children, and views about the relationship between the economy and family size.

Problems. If a person does not happen to think that economic conditions are bad (which the question imposes as an assumption) and/or that economic conditions of whatever kind have any implications for family size, but if that person happens to think having two children is a good target for a family, it is not easy to answer the question. Moreover, whether a person agrees or disagrees, it is hard to know what the respondent agreed or disagreed with.

The agree-disagree format appears to be a rather simple way to construct questions. In fact, to use this form to provide reliable, useful measures is not easy and requires a great deal of care and attention. Usually, researchers will have more reliable, valid, and interpretable data if they avoid the agree-disagree question form.

INCREASING THE VALIDITY OF FACTUAL REPORTING

When a researcher asks a factual question of a respondent, the goal is to have the respondent report with perfect accuracy; that is, give the same answer that the researcher would have given if the researcher had access to the information needed to answer the question. There is a rich methodological literature on the reporting of factual material. Reporting has been compared against records in a variety of areas, in particular, the reporting of economic and health events (see Cannell, Marquis, & Laurent, 1977, for a good summary. Also Edwards et al., 1994; and Edwards, Winn, & Collins, 1996).

Respondents answer many questions accurately. For example, more than 90% of overnight hospital stays within 6 months of an interview are reported (Cannell et al., 1977). How well people report, however, depends on both what they are being asked and how it is asked. There are four basic reasons why respondents report events with less than perfect accuracy:

1. They do not understand the question.
2. They do not know the answer.

3. They cannot recall it, although they do know it.
4. They do not want to report the answer in the interview context.

There are several steps that the researcher can take to combat each of these potential problems. These steps are reviewed next.

Understanding the Question

If respondents do not all have the same understanding of what the questions ask for, error is certain to result. As discussed earlier, when researchers are trying to count events that have complex definitions, such as burglaries or physician services, they have two options: (a) Provide definitions to all respondents; or (b) have respondents provide the information needed to classify their experiences into detailed, complex categories, and then have coders categorize answers.

Fowler (1992) has shown that people do answer questions that include ambiguous terms, producing quite distorted data. Researchers cannot assume that respondents will ask for clarification if they are not sure what a question means. To maximize the validity of factual survey data, an essential first step is to write questions that will be consistently understood by all respondents.

Lack of Knowledge

Lack of knowledge as a source of error is of two main types: (a) The chosen respondent does not know the answer to the question, but someone in the selected household does; or (b) no one in the selected household knows the answer. The solution in the first situation lies in choosing the right respondent, not question design. Most often, the problem is that one household respondent is asked to report information about other household members or the household as a whole. Solutions include the following:

- Identify and interview the household member who is best informed.
- Use data collection procedures that permit the respondent to consult with other household members.
- Eliminate proxy respondents; ask respondents to provide information only about themselves.

Sometimes a complex data collection strategy is called for. For example, the National Crime Survey conducted by the Bureau of the Census obtains reports of household crimes from a single household informant, but in addition asks each household adult directly about personal crimes such as robbery. If the basic interview is to be carried out in person, costs for interviews with other members of the household can be reduced if self-administered forms are left to be filled out by absent household members, or if secondary interviews are done by telephone. A variation is to ask the main respondent to report the desired information as fully as possible for all household members, then mail the respondent a summary for verification, permitting consultation with other family members.

When respondents are asked questions about themselves that they cannot answer, it is a question design problem. In theory, one could differentiate between information the respondent cannot recall and information the respondent never had at all. In either case, the problem for the researcher is to design questions that almost everyone can answer. Among the options available are the following:

- Change the question to ask for information that is less detailed or easier to recall.
- Help the respondent estimate the answer.
- Change or drop the objective.

It is not uncommon for questions to ask for answers in more detail than the research objectives require.

> The question asks respondents for the name of all the medications they take (a very hard question) when the objective is to find out who is taking medicine for hypertension (a much easier question).

> The question asks for income in an open-ended (and implicitly very detailed) way when getting an estimate of income in broad categories would satisfy the research objectives.

Recall follows some obvious principles: Small events that have less impact are more likely to be forgotten than more significant events; recent events are reported better than events that occurred in the more distant past (Cannell, Marquis, & Laurent, 1977). Sometimes it may be worthwhile to change question objectives to improve reporting by asking about events that are easier to recall. For example, although it may be desirable to have respondents report all the crimes that happened in the past year, there will be less reporting error if they are asked to report for only 6 months.

A comparatively new set of question design strategies has resulted from the growing involvement of cognitive psychologists in survey methods (Jabine, Straf, Tanur, & Tourangeau, 1984; Sirken et al., 1999; Schwartz & Sudman, 1996). Various strategies are being tried to help respondents recall events (e.g., by suggesting possible associations) or place events in time (e.g., by having respondents recall something that happened about a year before). For many survey tasks, studies have shown that respondents do not actually use recall to answer some questions; they estimate the answers (e.g., Burton & Blair, 1991). For example, if respondents are asked for the number of times they visited a grocery store to buy food in some period, they usually estimate based on their usual patterns rather than try to remember the individual events. This observation leads researchers to design strategies for helping respondents make better estimates.

Finally, it is important to recognize that there are some things that researchers would like to have people report that they cannot. For example, people do not know the cost of their medical care that is paid by insurance. If one truly wants to obtain medical costs, it is necessary to supplement what respondents can report (their out-of-pocket expenditures) with data collected directly from providers or insurers.

Social Desirability

There are certain facts or events that respondents would rather not report accurately in an interview. Health conditions that have some degree of social undesirability, such as mental illness and venereal disease, are underreported significantly more than other conditions. Hospitalizations associated with conditions that are particularly threatening, either because of the possible stigmas that may be attached to them or because of their life-threatening nature, are reported at a lower rate than average (Cannell, Marquis, & Laurent, 1977). Aggregate estimates of alcohol consumption strongly suggest underreporting, although the reporting problems may be a combination of recall difficulties and respondents' concerns about social norms regarding drinking. Arrest and bankruptcy are other events that have been found to be underreported consistently but seem unlikely to have been forgotten (Locander, Sudman, & Bradburn, 1976).

There are probably limits to what people will report in a standard interview setting. If a researcher realistically expects someone to admit something that is very embarrassing or illegal, extraordinary efforts are needed to convince respondents that the risks are minimal and that the reasons for taking any risk are substantial. The following are some of the steps that a researcher might consider when particularly sensitive questions are being asked (also see Catania, Gibson, Chitwood, & Coates, 1990; Sudman & Bradburn, 1982).

1. *Minimize a sense of judgment; maximize the importance of accuracy.* Careful attention to the introduction and vocabulary that might imply the researcher would value certain answers negatively is important.

Researchers always have to be aware of the fact that respondents are having a conversation with the researcher. The questions and the behavior of the interviewer, if there is one, constitute all the information the respondent has about the kind of interpretation the researcher will give to the answers. Therefore, the researcher needs to be very careful about the cues respondents are receiving about the context in which their answers will be interpreted.

2. *Use self-administered data collection procedures.* Although the data are not conclusive, there is some evidence that telephone interviews are more subject to social desirability bias than personal interviews (Aquilino, 1994; de Leeuw & van de Zouwen, 1988; Fowler, Roman, & Di, 1998; Henson, Roth, & Cannell, 1977; Mangione, Hingson, & Barret, 1982). The evidence is much clearer that having respondents answer questions in a self-administered form rather than having an interviewer ask the questions may produce less social desirability bias for some items (e.g., Aquilino, 1994, 1998; Aquilino & Losciuto, 1990; Dillman & Tarnai, 1991; Fowler, Roman, Di, 1998; Hochstim, 1967). For surveys dealing with sensitive topics, a mail survey or group administration should be considered. A personal interview survey also can include some self-administered questions: A respondent simply is given a set of questions to answer in a booklet. If the survey is computer assisted, the respondents can enter their answers directly into a computer with much the same effect. For example, such an approach has been shown to significantly increase reports of recent illegal drug use (Penne, Lessler, Beiler, & Caspar, 1998;

Tourangeau & Smith, 1998). Finally, Turner, Forsyth, and O'Reilly (1998) have shown that telephone surveys obtain much higher estimates of socially sensitive activities related to sex and drugs when answers are entered directly into a computer using the touch-tone feature on the telephone than when an interviewer asks the questions.

3. *Confidentiality and anonymity.* Almost all surveys promise respondents that answers will be treated confidentially and that no one outside the research staff will ever be able to associate individual respondents with their answers. Respondents usually are assured of such facts by interviewers in their introductions and in advance letters, if there are any; these may be reinforced by signed commitments from the researchers. Self-administered forms that have no identifiers provide a way to ensure that answers are anonymous—not just confidential. Finally, for surveys on particularly sensitive or personal subjects, there are some elaborate survey strategies, such as random response techniques, that ensure respondents cannot be linked to their answers (these are described by Fox & Tracy, 1986, and by Fowler, 1995).

Again, it is important to emphasize that the limit of survey research is what people are willing to tell researchers under the conditions of data collection designed by the researcher. There are some questions that probably cannot be asked of probability samples without extraordinary efforts. Some of the procedures discussed in this section, however, such as trying to create a neutral context for answers and emphasizing the importance of accuracy and the neutrality of the data collection process, are probably worthwhile procedures for the most innocuous of questions. Any question, no matter how innocent it may seem, may have an answer that is embarrassing to somebody in the sample. It is best to design all phases of a survey instrument with a sensitivity to reducing the effects of social desirability and embarrassment for any answers people may give.

INCREASING THE VALIDITY OF ANSWERS DESCRIBING SUBJECTIVE STATES

As discussed, the validity of subjective questions has a different meaning from that of objective questions. There is no external criterion; one can estimate the validity of a subjective measure only by the extent to which answers are associated in expected ways with the answers to other questions, or other characteristics of the individual to which they should be related (see Turner & Martin, 1984, for an extensive discussion of issues affecting the validity of subjective measures).

There basically are only three steps to the improvement of validity of subjective measures:

1. Make the questions as reliable as possible. Review the sections on the reliability of questions, dealing with ambiguity of wording, standardized presentation, and vagueness in response form, and do everything possible to get questions that will mean the same thing to all respondents. To the extent that subjective measures are unreliable, their validity will be reduced. A special issue is the reliability of ordinal scales, which are dominant among measures of subjective states. The response alternatives offered must be unidimensional (i.e., deal with only one issue) and monotonic (presented in order, without inversion).

PROBLEMATIC SCALES

5.21 How would you rate your job—very rewarding, rewarding but stressful, not very rewarding but not stressful, or not rewarding at all?

5.22 How would you rate your job—very rewarding, somewhat rewarding, rewarding, or not rewarding at all?

Question 5.21 has two scaled properties, rewardingness and stress, that need not be related. Not all the alternatives are played out. Question 5.21 should be made into two questions if rewardingness and stress of jobs are both to be measured. In 5.22, some would see "rewarding" as more positive than "somewhat rewarding" and be confused about how the categories were ordered. Both of these problems are common and should be avoided.

2. When putting people into ordered classes along a continuum, it probably is better to have more categories than fewer. There is a limit, however, to the precision of discrimination that respondents can exercise in giving ordered ratings. When the number of categories exceeds the respondents' ability to discriminate their feelings, numerous categories simply produce unreliable noise. Also, numerous categories may make questions harder to administer, particularly on the telephone. However, to the extent that real variation among respondents is being measured, more categories will increase validity.
3. Ask multiple questions, with different question forms, that measure the same subjective state; combine the answers into a scale. The answers to all questions potentially are influenced both by the subjective state to be mea-

sured and by specific features of the respondent or of the questions. Some respondents avoid extreme categories; some tend to agree more than disagree. Multiple questions help even out response idiosyncrasies and improve the validity of the measurement process (Cronbach, 1951; DeVellis, 1991).

The most important point to remember about the meaning of subjective measures is their relativity. Distributions can be compared only when the stimulus situation is the same. Small changes in wording, changing the number of alternatives offered, and even changing the position of a question in a questionnaire can make a major difference in how people answer (see Schuman & Presser, 1981; Sudman & Bradburn, 1982; and Turner & Martin, 1984, for numerous examples of factors that affect response distributions). The distribution of answers to a subjective question cannot be interpreted directly; it has meaning only when differences between samples exposed to the same questions are compared or when patterns of association among answers are studied.

QUESTION DESIGN AND ERROR

A defining property of social surveys is that answers to questions are used as measures. The extent to which those answers are good measures is obviously a critical dimension of the quality of survey estimates. Questions can be poor measures because they are unreliable (producing erratic results) or because they are biased, producing estimates that consistently err in one direction from the true value (as when drunk-driving arrests are underreported). We know quite a bit about how to make questions reliable. The principles outlined in this chapter to increase reliability are probably sound. Although other points might be added to the list, creating unambiguous questions that provide consistent measures across respondents is always a constructive step for good measurement.

The validity issue is more complex. In a sense, each variable to be measured requires research to identify the best set of questions to measure it and to produce estimates of how valid the resulting measure is. Many of the suggestions to improve reporting in this chapter emerged from a 20-year program to evaluate and improve the measurement of health-related variables (Cannell, Marquis, & Laurent, 1977; Cannell, Oksenberg, & Converse, 1977). There are many areas in which a great deal more work on validation is needed.

Reducing measurement error through better question design is one of the least costly ways to improve survey estimates. For any survey, it is important to attend to careful question design and pretesting (which are discussed in Chapter 6) and to make use of the existing research literature about how to measure what is to be measured. Also, continuing to build a literature in which the validity of measures has been evaluated and reported is much needed. Robinson, Shaver, and Wrightsman (1997) and McDowell and Newell (1996) have compiled data on the validity of many commonly used multi-item measures that document how measures have been validated, as well as how much work remains to be done.

EXERCISES

Use the criteria discussed in this chapter to evaluate the following questions as reliable, interpretable, and analytically useful measures; write better questions if you can.

a. To measure income: How much do you make?
b. To measure health: How healthy are you?
c. To measure satisfaction with life: How would you rate your life—very good, better than average, mixed, could be better, or very bad?
d. To measure opinion about abortion laws: Tell me whether you agree or disagree with the following statement: Abortion is morally very questionable; abortions should be illegal, except in emergencies.

Write a hypothesis about a possible relationship between two variables (e.g., good health is associated with receiving good-quality health care; or good-quality housing is related to having a high income). Describe the information you would need in order to assign a value to a person for each of the two variables. Then draft a question (or set of questions) to characterize respondents on each of the two variables specified in your hypothesis, the answers to which would provide the information you need. Indicate whether your questions ask for factual or subjective information and whether the resulting data will have nominal, ordinal, interval, or ratio properties.

6

Evaluating Survey Questions and Instruments

Designing a good survey instrument involves selecting the questions needed to meet the research objectives, testing them to make sure they can be asked and answered as planned, then putting them into a form to maximize the ease with which respondents and interviewers can do their jobs. This chapter describes steps for designing good survey instruments.

Every survey requires either an interview schedule, which constitutes a script for survey interviewers, or a questionnaire that respondents will read and fill out themselves. These documents, either in paper form or as programs for a computer, will be referred to generically as survey instruments.

Understanding what a good question is and how to use questions as measures, as discussed in Chapter 5, is certainly the foundation of good survey instrument design. There is, however, a series of very practical steps needed to produce a good data collection instrument. This chapter presents a summary of those steps. Sudman and Bradburn (1982), Converse and Presser (1986), Bradburn and Sudman (1992), and Fowler (1995) provide longer, more detailed discussions of such steps.

Survey instrument design has two components: deciding what to measure and designing and testing questions that will be good measures. The first step usually is to define the survey objectives, though those objectives may be revised based on subsequent question testing. Then the process of choosing and testing questions takes place. The steps involved in a survey instrument development process may include the following:

- focus group discussions
- drafting a tentative set of questions
- critical review to detect common flaws
- individual laboratory interviews (not replicating proposed data collection procedures)
- putting questions into a survey instrument
- pretesting using an approximation of proposed data collection procedures

DEFINING OBJECTIVES

A prerequisite to designing a good survey instrument is deciding what is to be measured. This may seem simple and self-evident, but it is a step that often is overlooked, to the detriment of the results. One valuable first step is to write a paragraph about what the survey is supposed to accomplish. In designing a survey instrument, researchers often are tempted to add related questions that do not contribute to achieving the project's goals. A check against such temptations is to have a good statement of the purposes, against which the inclusion of a particular area of inquiry can be measured. Second, one should make a list of what should be measured to accomplish the goals of the project. These should not be questions; they should be variables to be measured, listed in categories or areas that make sense.

An analysis plan should be developed to go with the list of variables to be measured. Presumably, a good start already will have been made in connection with the design of the sample. The researcher will have had to think through which subgroups in the population require special estimates. At this point, however, the researcher should refine those ideas so that there is a clear list of (a) which variables are designed to be dependent variables, for which measures of central tendency (e.g., means or distributions) are to be estimated; (b) which variables are needed as independent variables in order to understand distributions and patterns of association; and (c) which variables may be needed as control or intervening variables to explain patterns observed and to check out competing hypotheses.

These three documents, a statement of purposes, a list of the kinds of variables to be measured, and a draft of an analysis plan, are essential components for developing a survey instrument.

PRELIMINARY QUESTION DESIGN STEPS

Focus Groups

Before writing a draft of a structured set of questions, it almost always is valuable to conduct focused discussions with people who are in the study population about the issues to be studied. The primary purpose of these discussions is to compare the reality about which respondents will be answering questions with the abstract concepts embedded in the study objectives.

Example. The goal is to measure the number of visits to doctors. A group discussion could be focused on what counts as a visit to a doctor. Two key concepts are "visit" and "doctor." Participants could be asked about the various contacts they had related to doctors (e.g., telephone consultations, trips to have X rays or laboratory tests, inoculations) and whether or not they considered these contacts to be visits. They also could be asked about the various people they contacted related to their health (e.g., psychologists, psychiatrists, physician assistants, ophthalmologists, optometrists, physical therapists) and asked about whether or not they considered these individuals to be doctors.

This discussion alone could provide critical information of at least three types:

1. The kinds of contacts people have that possibly could be considered visits. This information would help the researcher refine the objectives and refine question wording to make it clear what is and is not to be included. For example, do we want to include telephone consultations? If a nurse practitioner is seen in a doctor's office, does that count?
2. What people know. For example, is everyone clear that a psychiatrist is an M.D., but a psychologist is not? What assumptions can be made about people's knowledge and perceptions of the background, training, or credentials of health care providers?
3. Comprehension of some key words or terms. Does the word *doctor* mean an M.D., or is it more generic (like Kleenex), referring to professionals in white coats delivering health-related services? Do alternative words, such as *health care provider* or *health care professional,* have consistent meaning for respondents?

Focus group discussions are best with six to eight people. The general protocol is to discuss people's perceptions, experiences, and perhaps feelings related to what is to be measured in the survey. The number of groups that is valuable will vary, but virtually every survey instrument will benefit from at least a couple of focus group discussions at an early stage in the survey instrument development process.

Drafting Questions

Armed with a list of what is to be measured, the researcher attempts to find the single question or set of questions needed to create measures of the variables on the list. Many questions, such as those dealing with background or demographic issues, are standard to many surveys. Reviewing the questions in the General Social Survey carried out by the National Opinion Research Center at the University of Chicago may be useful. Many surveys are also available online through the International Consortium of Political and Social Research (ICPSR) at the University of Michigan. Copies of original survey instruments from any of the major survey organizations also are useful as references. From these, the researcher can glean ideas about how specific questions are phrased, how to generate standardized questions, and how to format survey instruments.

Taking advantage of the work that others have done is very sensible. Of course, it is best to review questions asked by researchers who have done previous work on the study topic. In addition, if questions have been asked of other samples, collecting comparable data may add to the generalizability of the research. The mere fact that someone else has used a question before, however, is no guarantee that it is a very good question or, certainly, that it is an appropriate question for a given survey. Many bad questions are asked over and over again because researchers use them uncritically. All questions should be tested to make sure that they "work" for the populations, context, and goals of a particular study.

PRESURVEY EVALUATION

Critical Systematic Review

Once a set of questions is drafted, a good next step is to subject them to a critical systematic review. Lessler and Forsyth (1996) produced a list of issues to look for in a set of questions. Fowler (1995) also proposed a list of standards for questions that can be applied prior to testing. While neither list is exhaustive, both lists identify a set of question characteristics that are indicative of problem questions. Using one of these lists can help to identify questions that need revision; it also can flag issues for attention during the next phases of testing.

Cognitive Laboratory Interviews

Once a set of questions has been drafted, critically reviewed, and revised as warranted, the next step is to find out if they are questions people consistently can understand and can answer. Focus group discussions should provide some insights into comprehension issues, but they do not provide a forum for evaluating specific wording or the difficulty of the response task. At early stages of framing questions, the researcher also can learn a great deal by trying out questions on friends, relatives, and coworkers. Early versions of most survey instruments contain questions that are confusing, that cannot be read as written, and that are virtually unanswerable by anyone.

Once questions are in draft form, but before subjecting them to a formal field pretesting, a more formal kind of testing, commonly called cognitive testing, is a valuable next step (DeMaio & Rothgeb, 1996; Forsyth & Lessler, 1992; Fowler, 1995; Lessler & Tourangeau, 1989; Willis, DeMaio, & Harris-Kojetin, 1999). Although cognitive interviews take a variety of forms, there are certain features that they usually share. First, respondents are volunteers who are willing to spend more time than the data collection itself actually involves in order to help the researchers understand how the questions work. Often respondents are paid and are brought into a laboratory setting where the interviews can be videotaped.

These interviews usually are not done by regular interviewers. In some cases, interviewers are cognitive psychologists; in other cases, interviews are done by the investigators themselves or senior interviewer supervisors. In most cases, interviewers are thoroughly knowledgeable about the objectives of each question, so that they can detect issues that arise in the way that respondents understand questions or form answers to questions.

A typical protocol calls for asking respondents a set of proposed questions, then in some way gathering information about how the respondents understood the questions and about the way in which they answered them. Sometimes respondents are asked to "think aloud" while they are preparing their answers. In other cases, respondents are asked a set of questions about the way they understood each question and about issues related to their answers. Two of the most common tasks are

1. to ask respondents to say in their own words what they think the question is asking
2. to ask respondents to explain how they chose a particular answer over others

The point is to get enough information about the respondents' comprehension and preparation of responses to evaluate whether they performed the tasks in the way the researcher wanted. There are four specific kinds of questions that most cognitive testing is designed to answer:

1. Are questions consistently understood?
2. Do respondents have the information needed to answer the questions?
3. Do the answers accurately describe what respondents have to say?
4. Do the answers provide valid measures of what the question is designed to measure?

There are limits to what can be learned from laboratory interviews. Usually few such interviews are done (often fewer than 10), because they are labor-intensive and, in most organizations, can be conducted by only a small number of people. Second, the interviews are conducted under artificial conditions; tasks that volunteers are able and willing to perform may not be handled by a cross-section sample. Nonetheless, such interviews are increasingly seen as an essential step in the design and evaluation of a survey instrument. Questions that are not consistently understood or answered in a laboratory setting certainly will not work any better in an actual survey (Royston, 1989). Problems of comprehension and difficulties with the response task are not identified as reliably in field pretests as they are in laboratory interviews, where the answering process can be examined.

The cognitive laboratory interview has most often been used to test interview protocols. The same issues of comprehension and difficulty of the response task, however, apply to self-administered forms. Although standard tests of self-administered forms, as described next, often involve debriefing questions similar to those used in cognitive interviews, respondent comprehension is more apparent when the question-and-answer process is carried out orally. Thus, to test questions designed to be self-administered, an oral cognitive interview may be an effective way to identify problems that will not be picked up in the standard pretest.

DESIGN, FORMAT, AND LAYOUT OF SURVEY INSTRUMENTS

Once a set of questions is close to ready for final pretesting, the questions need to be put into a form to facilitate interviewer or self-administration. A first step is simply to order the questions. Many researchers like to start with relatively easy, straightforward questions that help get the respondent into the survey. Questions requiring a good deal of thought, or those believed to be sensitive, often are reserved for the middle or later sections of survey instruments. A good practical step is to number questions in sections: A1, A2, B1, B2, and so forth. In this way, when questions are added or deleted, it is not necessary to renumber every question.

Whether the survey is to be interviewer administered or self-administered, the goal of the layout and format of the questionnaire should be to make the tasks of the interviewer and the respondent as easy as possible. For an interviewer-administered survey instrument, the following are some rules that will help achieve that goal:

1. Adopt a convention that differentiates between the words that interviewers are to read to respondents and words that are instructions. A common convention is to use uppercase letters for instructions and lowercase for questions to be read aloud.
2. If an interview uses a paper-and-pencil form, and is not computer assisted, establish a clear convention for handling instructions to skip questions that do not apply to a particular respondent. The instructions should be keyed to a particular response and tell the interviewer where to go to ask the next questions. Of course, computer-assisted instruments will make skips automatically.
3. Put optional wording in parentheses. Conventions such as (his/her) or (husband/wife) are easy for interviewers to handle smoothly if they are alerted by the parentheses. A similar convention uses all caps (e.g., SPOUSE) when the interviewer must supply a word that is not provided in the question itself. Computer assistance often enables optional wording to be filled in, rather than have the interviewer adapt the wording to the situation.
4. Check to make sure that all the words that an interviewer has to say are, in fact, written. This includes not only the phrasing of the questions but transitions, introductions to questions, needed definitions, and explanations.

For self-administered questionnaires, the same kinds of general principles apply; that is, the main goal is to make the questionnaire easy to use. If anything, the formatting of a self-administered questionnaire is more important. In contrast to interviewers, respondents do not receive the benefit of training, they usually are not motivated to do the job well, and they are not selected on the basis of their ability to handle questionnaires. Five guiding principles are as follows:

1. A self-administered questionnaire mainly should be self-explanatory. Reading instructions should not be necessary, because they will not be read consistently.
2. Self-administered questionnaires mainly should be restricted to closed answers. Checking a box, clicking on a response, or circling a number should be the only tasks required. When respondents are asked to answer in their own words, the answers usually are incomplete, vague, and difficult to code, and therefore they are of only limited value as measurements.
3. The question forms in a self-administered questionnaire should be few in number. The more the instrument can be set up so that the respondent has the same kinds of tasks and questions to answer, the less likely it is that respondents will become confused; also, the easier the task will be for the respondents.
4. A questionnaire should be laid out in a way that seems clear and uncluttered. Photo reduction (or other strategies for putting many questions on a page) actually reduces the response rate compared with when the same number of questions are spaced attractively over more pages.
5. Provide redundant information to respondents by having written and visual cues that convey the same message about how to proceed. If people possibly can be confused about what they are supposed to do, they will be. Work on making everything simple and clear.

The design of computer-assisted instruments is still evolving. Many of the principles outlined here no doubt apply to computer-assisted instruments. However, Couper, Hansen, and Sadowsky (1997) have found that interviewers have significant problems with some complex CAPI programs, which suggests the need for further developmental work (also see

Couper, 1999). Dillman (2000) describes the results of his work on designing optimal Internet instruments, and that, too, is work in progress.

FIELD PRETESTS

Once a survey instrument has been designed that a researcher thinks is nearly ready to be used, a field pretest of the instrument and procedures should be done. The purpose of such pretests is to find out how the data collection protocols and the survey instruments work under realistic conditions.

Pretesting an Interview Schedule

The traditional pretest done by conscientious survey organizations usually consists of experienced interviewers taking 20 to 50 interviews with respondents drawn from a population the same as, or similar to, the population to be included in the survey. Interviewers are asked to play two roles in such pretests: They are interviewers, carrying out the procedures, and they are observers of the data collection process who are asked to report back to the researchers about any ways in which the procedures and survey instruments could be improved. It probably is most typical for this feedback process to take place in a group debriefing session, though on occasion interviewers report back individually.

Pretests such as those described are an essential part of the survey design process. A particularly important function is to test the usability of the instrument, both the questions and the layout, from the interviewers' perspective. However, such tests also have several limitations. The standards that interviewers use for what constitutes a problem often are not specified well, and it is almost certain that interviewers are inconsistent in what they consider to be a problem. In addition, a group discussion is an imperfect way to gather systematic information about the pretest experience.

Researchers have added steps designed to make the pretest experience more systematic and more valuable. One simple innovation is to ask interviewers to fill out a brief rating form on each question in addition to reporting back in a group session. One such rating form asks interviewers to evaluate each question with respect to whether or not (a) it is easy to read as worded, (b) respondents understand the question in a consistent way, and (c) respondents can answer the question accurately (Fowler, 1995). Obviously, interviewers have to guess about whether or not respondents are understanding questions and answering accurately; however, they do this in any case. The advantage of a form is that interviewers are asked systematically to attend to these aspects of question design as well as the other, more practical aspects of the survey instrument to which they ordinarily attend. Also, having interviewers do these ratings makes it easier for investigators to summarize interviewer reports and identify question problems in a more consistent way.

A more important, and probably more useful, innovation with respect to the field pretest is the use of tape recording and behavior coding to evaluate survey questions. With respondent permission, which is almost always granted, it is easy to tape-record pretest interviews done either in person or over the telephone. Trained coders can then listen to those tape recordings and evaluate problems in the question-and-answer process in a consistent way.

Three behaviors have been shown to be most important in identifying problems with survey questions (Fowler & Cannell, 1996; Oksenberg, Cannell, & Kalton, 1991): (a) whether or not the interviewer reads the question as worded, (b) whether or not the respondent asks for clarification, and (c) whether or not the respondent initially gives an inadequate answer that requires interviewer probing. It has been found that questions consistently produce or do not produce these kinds of behaviors in interviews; that is, there are questions that interviewers consistently misread, that lead respondents to ask for clarification, or that respondents consistently answer in an inadequate way. Such coding does not identify all questions that are not consistently understood by respondents. However, when one of these behaviors occurs in 15% or more of pretest interviews, it has been shown that a question is either highly likely to produce distorted data or distinctly susceptible to interviewer effects (Fowler, 1991; Fowler & Mangione, 1990).

An additional benefit of behavior coding of pretest interviews is that the results are systematic and can be replicated. Thus the question evaluation process is moved beyond the subjective opinions of researchers and interviewers, and concrete, replicable evidence is produced about questions that are inadequate.

Trace files are a third source of information from a pretest of a computer-assisted interview. When an interview is computer assisted, it is possible to retrieve the actual key strokes interviewers make. Those files can identify places where interviewers have to go back to previous screens and questions. Having to return to previous questions slows down an interview and often is a sign that question flow is not well designed. Looking at how "help" functions are used can provide clues to where help is needed and how "useful" various help functions are. Again, a plus of examining trace files is that the results are systematic and quantifiable (Couper, Hansen, & Sadowsky, 1997).

Pretesting a Self-Administered Questionnaire

If anything, self-administered instruments deserve more pretesting than interviewer-administered survey instruments, simply because interviewers can solve some problems that the researchers did not solve in the design of the survey instrument. Unfortunately, pretesting a self-administered instruments is also somewhat harder, because problems of comprehension and difficulties with answering questions are less evident. Although people have used observation of how people fill out forms or interact with a computer as a way of trying to identify unclear or confusing questions and instructions, it is not as satisfactory as the tape recording and behavior coding of interviews to identify question problems.

Probably the best way to pretest a self-administered questionnaire is in person, with a group of potential respondents. If it is a computer-based survey, respondents can use individual laptops. First, respondents should complete the questionnaire as they would if they were part of a survey. Then the researcher can lead a discussion about the instrument. One topic obviously is whether the instructions were clear. A second is whether or not the questions were clear. A third is whether there were any problems in understanding what kinds of answers were expected, or in providing answers to the questions as posed.

In addition to group tests, the usability of a computer-based instrument often benefits from some one-on-one testing, in which some respondents are observed interacting with the computer and the questions.

Debugging a Computer-Assisted Instrument

Having interviewers or respondents test instruments provides information about ease of use, but it does not provide information about whether or not the data collection protocol is correct. The key area for concern is the "skip" instructions.

A great strength of computer assistance is to help respondents and interviewers correctly navigate contingencies: When which questions are asked, or how they are worded, is tied to the answers given to preceding questions. Of course, the accuracy of the "skip" instructions requires careful proofreading of the final versions of paper instruments. However, the challenges of checking the accuracy of computer-assisted instruments are much greater than for paper instruments. The problem is that testers cannot see which questions are skipped and, hence, they may miss the fact that a question is skipped that should have been asked. Proofreading a printout of the program plus extensive testing are valuable steps. However, if an instrument is long and contains complex contingencies, those steps may be inadequate.

For this reason, once a survey begins, it should be standard practice to tabulate the distributions of answers to the early returns. It is only by checking such output that a researcher can be sure that the contingency instructions are working as intended.

SURVEY INSTRUMENT LENGTH

One outcome of a good pretest is finding out how long it takes to complete a survey instrument. The criteria for interview length should include cost, effect on response rate, and the limits of respondent ability and willingness to answer questions. The extent to which the length of a self-administered questionnaire affects costs and response rates varies with the population being studied and the topic; generalizations are difficult. It also is hard to generalize about how long people can be interviewed.

When researchers find they have more questions to ask than they feel they can ask, there are two choices available. Of course, the researcher simply may cut questions. An alternative approach is to assign subsets of questions to representative subsamples of respondents. Such an approach increases the complexity of the survey and reduces the precision of estimates of those variables, but this may be preferable to leaving out questions altogether. A clear advantage of computer-assisted data collection is the ease with which such designs can be implemented.

CONCLUSION

There was a time when one might have thought that evaluation of questions was largely a subjective process, contingent on the taste and preference of interviewers and researchers. We now know we can move beyond that. Survey questions should mean the same thing to all respondents; answering the questions should be a task that most or almost all respondents can perform; and the words in an interview schedule should be an adequate script that interviewers can follow as worded in order to conduct an interview.

Obviously, no matter how clear the question, some respondents will have difficulty with it, and some interviewers will misread it. There are judgments to be made about how bad a question can be before it must be changed. A critical part of the design and evaluation process of survey instruments, however, is to gather information about comprehension, the task of answering questions, and how interviewers and respondents use the protocols, in order that judgments can be made about whether or not the questions and instruments need to be changed. Good question and instrument evaluation prior to actually doing a survey is a critical part of good survey practice. It is one of the least expensive ways to reduce error in survey estimates. Although there is work to be done to define the most efficient and effective ways of evaluating questions, the procedures outlined on the preceding pages constitute a useful array of techniques that, when used, will have a major positive impact on the quality of survey data.

EXERCISE

Take the questions generated in the exercise for Chapter 5 and transform them into a set of questions that an interviewer could administer in a standardized way. Pretest and revise as needed. Now put the same questions in a form for self-administration. Pretest that.

7

Survey Interviewing

Interviewers affect survey estimates in three ways: They play a major role in the response rate that is achieved, they are responsible for training and motivating respondents, and they must handle their part of the interview interaction and question-and-answer process in a standardized, nonbiasing way. This chapter discusses the significance of interviewer selection, training, and supervision, plus the procedures interviewers are given, for minimizing interviewer-related error in surveys.

OVERVIEW OF INTERVIEWER JOB

Although some surveys are carried out using self-administered methods, using interviewers to ask questions and record answers is certainly a common part of survey measurement procedures, both face to face and over the telephone. Because of the central role they play in data collection, interviewers have a great deal of potential for influencing the quality of the data they collect. The management of interviewers is a difficult task, particularly in personal interviewer studies. The goal of this chapter is to provide an understanding of what an interviewer is supposed to do, appropriate procedures for managing interviewers, and the significance of interviewer management and performance for the quality of survey-based estimates.

Interviewers have three primary roles to play in the collection of survey data:

- to locate and enlist the cooperation of selected respondents
- to train and motivate respondents to do a good job of being a respondent
- to ask questions, record answers, and probe incomplete answers to ensure that answers meet the question objectives

Gaining Cooperation

Interviewers have to get in touch with respondents in order to enlist cooperation. The difficulty of this part of the job differs greatly with the sample. Interviewers have to be available when respondents want to be interviewed, they have to be available (and persistent enough) to make contact with hard-to-reach respondents, and for in-person interviews they have to be able and willing to go where the respondents are.

Although many sampled individuals agree readily to be interviewed, enlisting the cooperation of uninformed or initially reluctant respondents is undoubtedly one of the hardest and one of the most important tasks interviewers must perform. More interviewers probably fail in this area than in any other.

There is no doubt that some interviewers are much better than others at enlisting cooperation. It also is clear that different personal styles will work. Some effective interviewers are very businesslike, whereas others are more personable. Experience suggests that there are two characteristics that interviewers who are good at enlisting cooperation seem to share. First, they have a kind of confident assertiveness. They present the study as if there is no question that the respondent will want to cooperate. The tone and content of their conversation does not hint at doubt that an interview will result. Second, they have a knack for instantly engaging people personally, so that the interaction is focused on and tailored very individually to the respondent. It may be very task oriented, but it is responsive to the individual's needs, concerns, and situation. Reading a predesigned script is not an effective way to enlist cooperation.

Although these interviewer skills are important for all surveys, they are challenged particularly by telephone surveys for which respondents receive no advance notice (as in the case when random-digit dialing is used) or when the subject matter does not readily engage respondent interest.

Training and Motivating Respondents

Respondent performance, such as the accuracy of reporting, has been linked to their orientation to the interview. Interviewers have been shown to play an important role in setting respondent goals (Cannell & Fowler, 1964; Cannell, Oksenberg, & Converse, 1977; Fowler & Mangione, 1990). For example, interviewers who rush through interviews encourage respondents to answer questions quickly. Interviewers who read questions slowly indicate to respondents, in a nonverbal way, their willingness to take the time to obtain thoughtful, accurate answers; consequently, they do obtain more accurate answers. Studies also show that the way interviewers provide encouragement to respondents affects their sense of what they are supposed to do and how well they report (Cannell, Groves, Magilvey, et al., 1987; Cannell, Oksenberg, & Converse, 1977; Fowler & Mangione, 1990; Marquis, Cannell, & Laurent, 1972).

There is no doubt that most respondents have little idea of what they are expected to do and how they are to perform their roles. Interviewers both explicitly and implicitly teach respondents how to behave; this is an often unappreciated but critical part of the interviewer's job.

Being a Standardized Interviewer

Survey researchers would like to assume that differences in answers can be attributed to differences in what respondents have to say (i.e., their views and their experiences) rather than to differences in the stimulus to which they were exposed (i.e., the question wording, the context in which it was asked, and the way it was asked). The majority of interviewer training is aimed at teaching trainees to be standardized interviewers who do not affect the answers they obtain. There are five aspects of interviewer behavior that researchers attempt to standardize: the way they present the study and the task; the way questions are asked; the way inadequate answers (i.e., answers that do not meet question objectives) are probed; the way answers are recorded; and the way the interpersonal aspects of the interview are handled. Each of these is discussed next in greater detail.

1. *Presenting the study.* Respondents should have a common understanding of the purposes of the study, because this sense of purpose may have a bearing on the way they answer questions. Assumptions about such things as confidentiality, the voluntary nature of a project, and who will use the results also potentially can have some effect on answers. A good interviewing staff will give all respondents a similar orientation to the project so that the context of the interview is as constant as possible.

2. *Asking the questions.* Survey questions are supposed to be asked exactly the way they are written, with no variation or wording changes. Even small changes in the way questions are worded have been shown, in some instances, to have significant effects on the way questions are answered.

3. *Probing.* If a respondent does not answer a question fully, the interviewer must ask some kind of follow-up question to elicit a better answer; this is called probing. Interviewers are supposed to probe incomplete answers in nondirective ways—ways that do not increase the likelihood of any one answer over another. A short list of standard probes, including repeating the question, asking "Anything else?", "Tell me more?", and "How do you mean that?", will handle most situations if the survey instrument is designed well.

4. *Recording the answers.* The recording of answers should be standardized so that no interviewer-induced variation occurs at that stage. When an open-ended question is asked, interviewers are expected to record answers verbatim; that is, exactly in the words that the respondent uses, without paraphrasing, summarizing, or leaving anything out. In closed-response questions, when respondents are given a choice of answers, interviewers are required to record an answer only when the respondent actually chooses one. There is potential for inconsistency if interviewers code respondent words into categories that the respondent did not choose.

5. *Interpersonal relations.* The interpersonal aspects of an interview are to be managed in a standardized way. Inevitably, an interviewer brings some obvious demographic characteristics into an interview, such as gender, age, and education. By emphasizing the professional aspects of the interaction and focusing on the task, however, the personal side of the relationship can be minimized. Interviewers generally are instructed not to tell stories about themselves or to express views or opinions related to the subject matter of the interview. Interviewers are not to communicate any judgments on answers that respondents give. In short, behaviors that communicate the personal, idiosyncratic characteristics of the interviewer are to be avoided because they will vary across interviewers. To behave as a professional, not a friend, helps to standardize the relationship across interviewers and respondents. There is no evidence that having a friendly interpersonal style per se improves the accuracy of reporting; it probably tends to have a negative effect on accuracy (Fowler & Mangione, 1990).

A special complexity is introduced when the interviewer and respondent come from different backgrounds in society. In this instance, communication may not be as free and easy as when backgrounds are similar. There is some evidence that interviewers who take steps to ease communication in such situations (e.g., by introducing a bit of humor) may be able to produce a more effective interview (Fowler & Mangione, 1990). Efforts to relax the respondent, however, should not detract from a basically professional interaction, focused on good task performance.

Significance of Interviewer's Job

It should now be clear that interviewing is a difficult job. Moreover, failure to perform the job may produce three different kinds of error in survey data:

- Samples lose credibility and are likely to be biased if interviewers do not do a good job of enlisting respondent cooperation.
- The precision of survey estimates will be reduced, there will be more error around estimates, to the extent that interviewers are inconsistent in ways that influence the data.
- Answers may be systematically inaccurate or biased to the extent that interviewers fail to train and motivate respondents appropriately or fail to establish an appropriate interpersonal setting for reporting what is required.

Given all this potential to produce error, researchers should be motivated to use good interviewers. There are several avenues for affecting the quality of an interviewer's work: recruitment and selection, training, supervision, designing good questions, and using effective procedures. The next five sections will discuss the potential of each of these to influence interviewer performance.

INTERVIEWER RECRUITMENT AND SELECTION

Some of the characteristics of interviewers are dictated by requirements of the survey interviewer's job that have nothing to do with the quality of data per se:

1. Interviewers must have reasonably good reading and writing skills. Many, if not most, interviewers now work with computers, so that typing skills and general familiarity with computers are pluses. Most survey research organizations require high school graduation, and many require or prefer interviewers to have at least some college experience.

2. Interviewing is primarily part-time work. It is difficult to work 40 hours a week every week on general population surveys; survey organizations almost always have some ebbs and flows of work for interviewers. As a result, potential interviewers usually are people who can tolerate intermittent income or are between more permanent jobs. Interviewer pay is usually not high for a college-educated person. Often, there are no bene-

fits, such as health insurance, to the interviewer job. It is unusual for a survey interviewer to be able to rely on interviewing as a sole source of income and support over a long period of time.

3. Personal household interviewers must have some flexibility of hours; surveys require interviewers to be available when respondents are available. One advantage of telephone interviewing is that individual interviewers can work more predictable shifts, although evening and weekend work is prime time for almost all general-population survey work.

4. Personal household interviewers must be mobile, which often excludes people with some physical disabilities and those without the use of a car. Neither of these restrictions is salient to telephone interviewers.

Beyond these practical job requirements, there is little research basis for preferring one set of interviewer candidates over others. For example, experienced interviewers are likely to be better at enlisting cooperation simply because those for whom it is a problem will not continue to work as interviewers; however, there is no documented positive effect of experience on data quality. There is some evidence that interviewers become careless and collect poorer data over time (Bradburn, Sudman, & Associates, 1979; Cannell, Marquis, & Laurent, 1977; Fowler & Mangione, 1990).

Likewise, having interviewers who have specialized knowledge about the subject matter is seldom a plus. In fact, because knowledgeable interviewers may assume they know what the respondent is saying when the respondent has not been clear, they may read more into what the individual is saying than people not trained in the area. Unless interviewer observations or ratings requiring an extensive specialized background are needed, a trained interviewer with no special background usually is the best choice.

Age, education, and gender of interviewer seldom have been associated with data quality, though there is some evidence that females may, on average, be more positively rated by cross-section samples (Fowler & Mangione, 1990; Groves, 1989). In general, a researcher would be best advised to send the best interviewer available to interview a respondent, regardless of demographic characteristics. The exception is if the subject matter of the survey directly bears on race or religion (or any demographic characteristic) and the feelings of the respondents about people in the same or different groups. For example, if people are to be interviewed

about their own anti-Semitic feelings, the Jewishness of the interviewer will make a difference in the answers (Robinson & Rhode, 1946). In the same way, blacks and whites express different feelings about race depending on the interviewer's skin color (Schuman & Converse, 1971).

It is important to note, however, that matching on ethnicity does not necessarily improve reporting. Two studies of this issue found that black respondents reported income from welfare (Weiss, 1968) and voting (Anderson, Silver, & Abramson, 1988) more accurately to white interviewers than to black interviewers.

There is no question that a researcher should consider the interaction between the subject matter of a survey and the demographic characteristics of the interviewers and respondents. If ethnicity (or some other characteristic) is extremely salient to the answers to be given, controlling the relationship of interviewer and respondent characteristics should be considered so that the effect of the interviewer on the data can be measured (Groves, 1989). For most surveys, however, the practical difficulties and costs of controlling interviewer assignments and the lack of predictable effects will argue against trying to control the demographic characteristics of respondents and interviewers.

Finally, volunteer interviewing staffs are almost always unsuccessful at carrying out probability sample surveys. There are several reasons for the failure of volunteers. Because it is hard to require attendance at lengthy training sessions, volunteers usually are trained poorly. Because it is hard to terminate poor volunteer interviewers, response rates are usually low. Moreover, volunteer attrition is usually high.

This discussion offers few guidelines for researchers in the selection of interviewers. In some rather specialized circumstances, the interviewer's ethnic background, age, or gender may affect answers; for example, teenagers may respond differently to older female interviewers (Erlich & Riesman, 1961). For most surveys, however, the particular job requirements largely will dictate the pool of interviewers. There is little basis for ruling out people because of their background or personality characteristics. Rather, the key to building a good interviewing staff is good training and careful supervision. In addition, because of the difficulty of identifying good interviewers in advance, attrition of less able interviewers is probably a critical and necessary part of building a good staff of interviewers.

TRAINING INTERVIEWERS

There is great diversity in the kinds of training experiences to which survey interviewers are exposed. The exact amount of time that will be devoted to training, the kind of training session, and the content of the program obviously will depend on the particular organizational setting and what interviewers are going to be doing. There is some disagreement, in addition, on the extent to which effort should be devoted to an initial training session, prior to the onset of field experience, versus continuous learning and retraining after interviewers have begun. Nonetheless, all professional survey organizations concerned about data quality have at least some kind of (usually face-to-face) training of all new interviewers. The following is a general summary of what reasonable interviewer training might entail.

Content of Training

The content of training includes both general information about interviewing that applies to all surveys and information specific to the particular study on which interviewers are to work. The general topics to be covered will include the following:

- procedures for contacting respondents and introducing the study
- the conventions that are used in the design of the survey instrument with respect to wording and skip instructions, so that interviewers can ask the questions in a consistent and standardized way
- procedures for probing inadequate answers in a nondirective way
- procedures for recording answers to open-ended and closed questions
- rules and guidelines for handling the interpersonal aspects of the interview in a nonbiasing way
- how to use the computer-assisted interviewing programs

In addition, many research organizations feel that it is a good idea to give interviewers a sense of the way that interviewing fits into the total research process. For that reason, they often attempt to give interviewers some familiarity with sampling procedures, coding, and the kinds of analyses and reports that result from surveys. Such information may be helpful to interviewers in answering respondent questions and may play a positive role in motivating the interviewer and helping him or her to understand the job.

With respect to any specific project, interviewers also need to know the following:

- Specific purposes of the project, including the sponsorship, the general research goals, and anticipated uses of the research. This information is basic to providing respondents with appropriate answers to questions and helping to enlist cooperation.
- The specific approach that was used for sampling, again to provide a basis for answering respondent questions. In addition, there may be some training required in how to implement the basic sample design.
- Details regarding the purposes of specific questions—not necessarily their roles in analyses, but at least the kind of information they are designed to elicit.
- The specific steps that will be taken with respect to confidentiality, and the kinds of assurances that are appropriate to give to respondents.

Procedures for Training

There are six basic ways to teach interviewers: written materials, lectures and presentations, computer-based tutorials, planned exercises, practice role-playing, and observation of early interviews. Written materials are usually of two types. First, it is a very good idea to have a general interviewer manual that provides a complete written description of interviewing procedures. In addition, for each particular study, there normally should be some project-specific instructions in writing. It is tempting when interviewers are being trained in person and a project is being done in a local site to skimp on the preparation of written materials. Newly trained interviewers, however, say that there is an overwhelming amount of material and information to absorb during training. Having the procedures in writing enables interviewers to review material at a more leisurely pace; it also increases the odds that messages are stated clearly and accurately.

Lectures and demonstrations obviously have a role to play in any interviewer training, whether only a single interviewer or a large group of interviewers is being trained. In addition to the general presentation of required procedures and skills, most trainers find that demonstrating a standardized interview is a quick and efficient way to give interviewers a sense of how to administer an interview. Videotapes are often used to supplement lectures. Videotapes of practice interviews or other interviewer activities are good tools for interviewer training. There are also some general training tapes that can be bought for use in interviewer training.

The widespread use of computer-assisted interviewing means that interviewer training must include teaching interviewers to use computer-based instruments. The most widely used survey systems have computer-based tutorials that can be integrated into general interviewer training.

Because these are new skills, supervised structured practice is one of the most important parts of interviewer training. Having interviewers take turns playing the respondent and interviewer roles is common practice. Practice should include enlisting cooperation and handling the question-and-answer process. There also is great value in monitoring some practice interviews with respondents who are not role-playing and whom interviewers do not know. For personal interviews, supervisors can accompany and observe new interviewers doing practice interviews or review tape-recorded interviews. On the telephone, interviews may be monitored directly or tape-recorded for later review.

Two studies (Billiet & Loosveldt, 1988; Fowler & Mangione, 1990) concluded that interviewer training of less than 1 day produces unsatisfactory interviewers; interviewers are not able to perform their jobs as instructed, and the resulting data are affected adversely. Training programs lasting from 2 to 5 days are the norm in professional survey organizations. The length of training depends on numerous factors, including the number of interviewers to be trained and the complexity of the project for which they are being trained. The critical key to the quality of training, however, is probably the amount of supervised practice interviewing.

SUPERVISION

The keys to good supervision are to have the information needed to evaluate interviewer performance and to invest the time and resources required to evaluate the information and provide timely feedback. There are four main aspects of interviewer performance to supervise: costs, rate of response, quality of completed questionnaires, and quality of interviewing. It is considerably easier to supervise interviewers who are doing telephone interviewing from a centralized facility than those interviewing in the field.

Costs

Supervising costs for interviewers requires timely information about time spent, productivity (usually interviews completed), and mileage charges for interviewers using cars. High-cost telephone interviewers are likely to be those who work at less productive times, have high refusal rates (a refusal takes almost as much time as an interview), or who simply find ways (e.g., editing interviews, sharpening pencils) to make fewer calls per hour. High-cost personal household interviewers are likely to live far from their sample addresses, to make trips that are too short or at the wrong times (evenings and weekends are clearly the most productive), or to have poor response rates.

Response Rates

It is critical to monitor response rates (particularly rates of refusals) by interviewers on a timely basis; however, this is not easy to do. There are three main problems:

1. For personal interviews, but not telephone surveys from a computerized central facility, it can be hard to maintain timely information about interviewer results.
2. Interviewers can understate their refusals by assigning unsuccessful results to other categories.
3. Assignments to in-person interviewers may not be comparable, so that differences in rates of refusals per interviewer may not be consistent indicators of interviewer performance. This issue applies much less to telephone interviewers working in centralized facilities.

Response rates cannot be calculated accurately until a study is over, but special efforts to identify refusals by the interviewer during data collection can alert supervisors to problems and are a very important part of interviewer supervision. It is difficult to help an interviewer who has response rate problems. On telephone studies, a supervisor can listen to introductions and provide feedback immediately after the interview (or noninterview) about how the interviewer might be more effective. For in-person interviewers, the task is more difficult because the supervisor cannot observe the interviewer's approach unless the supervisor accompanies the interviewer on a trip. Thus the supervisor often must be content with listening to the interviewer give a sample introduction.

Supervisors can give helpful hints to interviewers. It is important to make sure interviewers are fully informed about a survey. Having interviewers practice giving concise, clear answers to common questions may be useful. In addition to working on the details of introductions, supervisors may need to address an interviewer's general feeling about approaching people or about the survey project and its value. There are limits, however, to how much retraining will help; there are people who never can attain good response rates. Although it is stressful, one of the most effective ways to keep response rates high is to take ineffective interviewers off a study.

Review of Completed Survey Instruments

When interviewers are using paper-and-pencil instruments, a sample of completed survey instruments should be reviewed to assess the quality of data interviewers are collecting. When reviewing a completed interview, one obviously can look for whether the recording is legible, the skip instructions are followed appropriately, and the answers obtained are complete enough to permit coding. In addition, looking at a completed interview can give a pretty good idea of the extent to which an interviewer is recording respondent answers verbatim, as compared to recording summaries or paraphrases. For computer-assisted interviews, these issues—except for the recording and probing associated with narrative answers—are not relevant.

The Question-and-Answer Process

The quality of interviewing cannot be supervised by reviewing completed survey instruments; they do not tell the supervisor anything at all about the way the interviewer conducted the interview and how those answers were obtained. In order to learn this, a supervisor must directly observe the interviewing process.

A telephone survey from a central facility permits direct supervision of how the interviewer collects the data. A supervisor can and should be available to monitor interviewers at all times. Supervisors should listen systematically to all or parts of a sample of the interviews that each interviewer takes, evaluating (among other things) appropriate introduction of the study, asking questions exactly as written, probing appropriately and nondirectively, and appropriate handling of the interpersonal aspects of the interview. This process works best if a rating form covering these and other aspects of an interviewer's work is completed routinely by a monitor (Cannell & Oksenberg, 1988).

When interviewers are doing studies in respondents' homes or in other distant places, it is more difficult to supervise the question-and-answer process. There are only two ways to do it: A supervisor can accompany an interviewer as an observer, or interviews can be tape-recorded. Without tape recording or a program of observation, the researcher has no way to evaluate the quality of interviewing. All the most important aspects of the measurement process are unmonitored. Poor interviewers cannot be identified for retraining, and the researcher cannot report the quality of interviewing beyond saying that the interviewers were told what to do. Indeed, from the interviewer's point of view, it must be difficult to believe that standardized interviewing is important when it is the focus of training but is not attended to further. Fowler and Mangione (1990) present evidence that personal interviewers are less likely to interview the way they are trained if their work is not monitored directly by tape recording. Both Fowler and Mangione (1990), and Billet and Loosveldt (1988) found that data quality was improved when interviewers were monitored directly in this manner. It is now clear that direct supervision of the interview process should be a part of a well-managed survey.

SURVEY QUESTIONS

Although training and supervision are important to producing good interviewing, perhaps the most important step a researcher can take to produce good interviewing is to design a good survey instrument. Research has shown that certain questions are misread consistently, whereas others consistently are answered inadequately, requiring interviewers to probe to obtain adequate answers (Fowler, 1991; Fowler & Cannell, 1996; Fowler & Mangione, 1990; Oksenberg et al., 1991). These questions can be identified with the kind of pretesting described in Chapter 6.

The more interviewers have to probe, explain, or clarify, the more likely they are to influence answers. The better the survey instrument, the more likely it is that the interviewer will conduct a good, standardized interview. The role of good question design in producing good interviewing is discussed in detail in Fowler and Mangione (1990) and Fowler (1991).

INTERVIEWING PROCEDURES

Training and Motivating Respondents

Studies have demonstrated the value of going beyond good question design to help standardize the interview (Cannell et al., 1987; Cannell, Oksenberg, & Converse, 1977; Miller & Cannell, 1977). For example, the researcher can help the interviewer train the respondent in a consistent way. Before the interview begins, the interviewer might read something like the following:

> Before we start, let me tell you a little bit about the interview process, since most people have not been in a survey like this before. You will be asked two kinds of questions in this survey. In some cases, I will be asking you to answer questions in your own words. In those cases, I will have to write down every word you say, not summarizing anything. For other questions, you will be given a set of answers, and you will be asked to choose the one that is closest to your own view. Even though none of the answers may fit your ideas exactly, choosing the response closest to your views will enable us to compare your answers more easily with those of other people.

Interestingly, interviewers like this instruction a great deal. It explains the respondents' task to them, and it makes the question-and-answer process go more smoothly. In fact, good interviewers give instructions such as these on their own. The value of providing explicit instructions is that it reduces differences among interviewers by having them all do the same thing. In addition, such instructions have a salutary effect on the interviewer's performance. Once the interviewer has read an instruction explaining the job expectations, it is easier to do the job the way it should be done, and it is a little harder to do it wrong, because the respondent now also knows what the interviewer is supposed to do (Fowler & Mangione, 1990).

Standardized instructions to respondents also can be used to set goals and standards for performance:

> It is very important that you answer as accurately as you can. Take your time. Consult records if you want. Ask me to clarify if you have any question about what is wanted.

Such statements ensure that respondents have a common understanding of their priorities. Some interviewers unintentionally promise respondents they will make it easy on respondents if the latter will just give the interview; interviewers who hurry communicate that speed is more important than accuracy. When an instruction such as the one just discussed is read, it forces accuracy and data quality to be a central part of the role expectations for both respondent and interviewer. One more source of between-interviewer variability is reduced, and the odds of good performance by both are increased.

Cannell, Oksenberg, and Converse (1977) report on an even stronger approach, requiring respondents to sign a form committing themselves to try their best to give accurate and complete information before they were allowed to be interviewed. Numerous refusals were expected but did not occur. Response rates were unaffected by the form, whereas reporting was improved. Oral strategies for accomplishing the same thing on the telephone have also proven to be effective in improving reporting (Cannell et al., 1987).

Cannell also has tried to standardize the reinforcement interviewers give to respondents. Interviewers often inadvertently reinforce undesirable respondent behaviors (e.g., quick, thoughtless answers; Marquis, Cannell, & Laurent, 1972). Cannell, Oksenberg, and Converse (1977) report that when interview schedules were designed that forced interviewers to praise good behavior (e.g., checking records or answering slowly), respondent reporting improved. Using such procedures is somewhat difficult on a routine basis, but the work emphasizes the need to minimize inappropriate reinforcement by interviewers.

In conclusion, there are critical parts of the interviewer's job besides the direct question-and-answer process. In particular, the interviewer is responsible for communicating to the respondent how the interview is to proceed: what the respondent is supposed to do, what the interviewer is going to do, and what their joint goals are. This aspect of the interviewer's job mainly has been left up to the interviewer, and not surprisingly, interviewers differ in how they do it in ways that affect data. By developing standardized instruction programs for respondents, researchers can make the job of the interviewer easier, reduce an important source of between-interviewer variance, and improve the extent to which interviewers and respondents behave in ways that will make the measurement process go better.

Standardized Wording

It was stated previously that asking questions exactly as worded is a foundation of standardized measurement, but not everyone agrees (Tanur, 1991). Critics of standardized interviewing have observed that some questions are not consistently understood by all respondents. When that is the case, they argue that it would produce better data if interviewers were free to clarify or explain the meaning of the question (e.g., Schober & Conrad, 1997). In a similar vein, critics note that some data collection tasks—for example, when the same information is being gathered about several different people or events—produce very stilted or awkward interactions when interviewers try to use only prescripted wording. In these instances, it is argued that giving interviewers more flexibility with wording would result in a more comfortable interviewer-respondent interaction (Schaeffer, 1992).

Some of the criticism of standardized interviewing is primarily the result of poorly designed questions (see Suchman & Jordan, 1990). When questions are unclear or provide awkward scripts for interviewers, the solution often is to write better questions, not to have interviewers redesign the questions (Beatty, 1995). There is real basis for concern that when interviewers are given flexibility to reword or explain the questions, they will do it in a way that changes the meanings of questions and makes the resulting data worse, not better (Fowler & Mangione, 1990). However, there are certain questions—such as repetitive series or when a few respondents need detailed definitions that would be cumbersome to provide to all respondents—that might be better handled by giving interviewers more flexibility. Moreover, when interviewers make changes in question wording, it has not consistently been shown to increase interviewer-related error or response error (Dykema, Lepkowski, & Blixt, 1997; Fowler & Mangione, 1990).

There have been some experiments giving interviewers more discretion about how to ask and probe questions (Conrad & Schober, 2000; Schober & Conrad, 1997). To date, the results have been mixed: The accuracy of some reports may be improved, but considerably increased interviewer training and sometimes longer interviews are involved. When and how to give interviewers more flexibility is a topic that warrants further experimentation. Meanwhile, for most surveys, designing questions that interviewers can and will ask exactly as worded remains the primary way to conduct a good survey.

VALIDATION OF INTERVIEWS

The possibility that an interviewer will make up an interview is a potential concern. The likelihood of this happening varies with the sample, the interviewing staff, and the field procedures. For the most part, concern

about validation is restricted to surveys in which interviewers are conducting interviews in respondents' homes or are doing telephone interviews from their own homes. In such cases, the actual collection of data is not observable by supervisors. The number of hours required to carry out an interview can be sufficient to motivate an interviewer to make up an interview rather than take the time and effort to carry it out.

In the long run, probably the best protection against faked interviews is to have a set of interviewers that has some commitment to the quality of the research and the organization. Such problems seem to occur most often with newly hired interviewers. Even organizations with an experienced, professional staff, however, routinely check a sample of interviews to make sure they actually were taken.

There are two approaches to this type of validation. One approach is to mail all respondents a brief, follow-up questionnaire asking about reactions to the interview. Probably a more common procedure is to have interviewers obtain a telephone number from every respondent; a sample is called by a supervisor. Simply knowing in advance that a validation by mail or telephone will be done is likely to be a deterrent to interviewer cheating. In addition, to be able to say that such a check was done may be reassuring to users of the data.

THE ROLE OF INTERVIEWING IN SURVEY ERROR

As noted at the onset of this chapter, interviewers affect response rates, the accuracy of reporting, and the consistency or precision of measurement. Each of these has a central role in the quality of a survey estimate.

One of the most observable effects of good survey management is the response rate. Although this issue is discussed more thoroughly in Chapter 3, it is worth repeating that the quality of an interviewing staff is critical to the rate of response that will be obtained in any particular survey.

It is more difficult to measure the error introduced by interviewers in the question-and-answer process. Often, survey error is undetectable. When asking questions about subjective states, objective checks for bias or inaccuracy are generally not meaningful, as was discussed in Chapter 5. There have been studies, however, in which researchers had objective measures of facts respondents were asked to report, permitting evaluation of the accuracy of reporting. In one such study (Cannell, Marquis, & Laurent, 1977), samples of households in which someone had been hospitalized in the year preceding were interviewed. The accuracy of reporting could be evaluated by comparing the health interview reports of hospital stays with hospital records. One measure of reporting accuracy was simply the percentage of known hospitalizations that was reported.

In this study, it was found that the number of interviews assigned to an interviewer correlated very highly ($r = .72$) with the percentage of hospitalizations that were unreported in the interview. Interviewers who had large assignments, with whatever pressures that were brought to bear on them, collected much less accurate data than those with small assignments.

A different study using the same criterion (the percentage of hospitalizations reported; Cannell & Fowler, 1964) reached a similar conclusion. In this case, half of an interviewer's respondents reported hospitalizations in an interview, whereas the other half completed a self-administered form regarding hospitalizations after the interviewer had completed the rest of the health interview. It was found that interviewers whose respondents reported with great accuracy when asked to report hospitalizations in the interview also had respondents who reported very well in the self-administered form after the interviewer had left ($r = .65$). This study suggested not only that interviewers had a critical role to play in affecting the error of their respondents' reporting, but also that one way in which interviewers affected respondent performance was the degree to which they motivated respondents to perform well. In both cases, the effect of the interviewer on reporting accuracy was clear.

In the absence of validating data, one cannot assess accuracy. However, it also is possible to assess the extent to which interviewers influence the answers of their respondents. If an interviewing staff were operating in a perfectly standardized way, one would be unable to explain any variation in answers by knowing who the interviewer was. To the extent that answers are predictable, in part, from knowing who did the interview, it can be concluded that the interviewer is inappropriately influencing answers. Groves (1989) thoroughly discusses the techniques for calculating the extent to which interviewers were affecting the answers to questions and summarizes the results of numerous studies in which interviewer effects were calculated. It turns out that for many questions that interviewers ask, one cannot see any effect of the interviewer on the answers. For between one third and one half of the questions in most surveys, however, interviewers significantly affect the answers.

The result of these interviewer effects is to increase the standard errors around survey estimates. The size of the multiplier depends on the size of

the intraclass correlation (rho) and on the average size of interviewers' assignments (see Groves, 1989; Kish, 1962). If the intraclass correlation is .01 (which Groves found to be about the average), and the average number of interviews per interviewer is about 31, the standard errors of means will be increased by 14% over those estimated from the sample design alone. When interviewer assignments average closer to 50, for items with an intraclass correlation of .02, the estimates of standard errors will be increased by 41%.

Out of this discussion there are several points to be made about the role of the interviewer in the total error structure of survey data:

1. In addition to their role in response rates, interviewers can be associated with the extent to which respondents give inaccurate answers in surveys and with measurement inconsistency. Existing evidence clearly indicates that interviewers are a significant source of error for many kinds of measures.
2. The training and supervision that interviewers receive can significantly increase the consistency of interviewers, thereby improving the reliability of estimates, and reduce bias. In particular, interviewers who receive minimal training (e.g., less than 1 day) and interviewers who receive minimal or no feedback about the quality of their interviewing are poorer interviewers.
3. Procedures that structure the training and instruction of respondents, minimize inappropriate interviewer feedback, and in general, control more of the interviewer's behavior can reduce interviewer effects on data and increase overall accuracy.
4. Better question design is a key to better interviewing.
5. One design option that has been unappreciated is the size of the average interviewer assignment. Although training and management costs may be lower if a smaller number of interviewers is used, researchers may pay a price in data reliability for allowing individual interviewers to take large numbers of interviews. Reducing average interviewer assignments often is a cost-effective way to increase the precision of survey estimates.
6. Virtually all reports of the reliability of survey estimates ignore the effects of interviewers on data. In part, this is because researchers cannot sort out interviewer effects from sampling effects when interviewers are assigned samples on a nonrandom basis, such as convenience or geographic proximity. Interviewer effects are a significant source of error, however, for many items in most surveys. Any report of the precision of a survey estimate that ignores interviewer effects is likely to be an underestimate of survey error.

In conclusion, the role of the interviewer in contributing to error in survey data has not been appreciated generally. Although most survey researchers know that some training is necessary for interviewers, procedures for training and supervising interviewers vary widely and often are not adequate. It is unusual for researchers to make any efforts beyond training and supervision to minimize interviewer effects. Yet, these aspects of survey design constitute some of the most cost-effective ways to improve the quality of survey data. The impact of the interviewer on survey estimates deserves a central place in the design and reporting of survey studies that it has not yet achieved.

EXERCISE

Tape-record some role-played interviews in which you and/or others use a standardized interview schedule (the questions developed in Chapter 6, or a schedule from another source). Then listen to the tapes and systematically evaluate interviewer performance by noting for each question at least the following errors: did not read question exactly as worded; probed an inadequate answer in a biasing (directive) way; failed to probe an unclear answer; or any other possibly biasing or unstandardized interpersonal behavior. The evaluations are particularly instructive if done by a group, so that interviewer errors can be discussed.

8

Preparing Survey Data for Analysis

Survey answers usually are transformed into data files for computer analysis. This chapter describes options and good practice for data formats, code development, coding procedures and management, data entry, and data checking procedures.

Once data have been collected by a survey, no matter what the methods, they almost invariably must be translated into a form appropriate for analysis by computer. This chapter is about the process of taking completed questionnaires and survey interviews and putting them into a form that can be read and processed by a computer. The process of coding or data reduction involves five separate phases:

- deciding on a format (the way the data will be organized in a file)
- designing the code (the rules by which a respondent's answers will be assigned values that can be processed by machine)
- coding (the process of turning responses into standard categories)
- data entry (putting the data into computer-readable form)

- data cleaning (doing a final check on the data file for accuracy, completeness, and consistency prior to the onset of analysis)

There are two kinds of errors that can occur in going from an answer to an entry in a data file. First, there can be transcription errors any time someone records an answer or number. Second, there can be coding decision errors, misapplications of the rules for equating answers and code values. The options for quality control are tied to the particular data entry and coding procedures chosen. Those options and various alternative procedures are discussed next.

FORMATTING A DATA FILE

Each analytic software package has its own conventions regarding how data should be formatted. The most important step before beginning to design a data entry process is to determine what programs will be used to analyze data and which specific conventions regarding file formats and missing data can be handled by those programs. The term *record* as used here refers to all the data that pertain to a single individual, case, or interview. A record can consist of one or more lines. Historically, an 80-column card was the unit onto which data were punched that corresponded to a line of data. Now data typically are put directly onto a hard disk or floppy disk, but they still are usually stored as an ordered set of lines for each record or interview. Although conventions and rules vary with facilities and the programs to be used, the following are some common issues:

1. Even though actual data cards are not used, a card-and-column format for specifying the location of data is still very common. In this format, after every 80 columns are filled, a new line is started; however, many analysis programs comfortably handle lines longer than the traditional 80 columns of data. If one of these software packages is used, there is no need to be constrained by a card-and-column format.
2. A serial identifier for each respondent usually goes in the same location on each line of data for a particular record or interview. It also helps in checking for the completeness of data files to have a line or card number in the same location of each line of data. These markers preserve the order of the data if they are sorted and are critical for checking files for completeness.
3. It eases coding, data entry, and programming tasks if the data are coded in the order that they appear in the survey instrument. This will reduce errors at these stages and represents a relatively cost-free means of quality control.

4. Multiple codes in a single field or column are acceptable to some computer programs, not to others. It probably is best to put a single positive entry in each field that contains data. Similarly, some computer programs interpret blanks as zeros, whereas others do not. If zero is meant, it is best actually to code a zero rather than leave a blank field; if the intention is to code for nonresponse, some specific value (usually numeric but sometimes blank) should be used.

CONSTRUCTING A CODE

A code is a set of rules that translates answers into numbers and vice versa (some systems accept alphabetic values, but the vast majority of codes in surveys use numeric codes only). Which numbers go with which answers is irrelevant to the computer. It is critical to reliable coding and appropriate interpretation of data, however, that the code be unambiguous. There should be a clear rule for what number to assign to each and every answer (or other result). In addition, codes can be designed to minimize errors during coding and analysis. The following are some common principles:

1. Be sure to have missing data codes for questions that are not answered. Codes should differentiate between the following:
   a. *Not ascertained information*, where codable information was not obtained as a result of imperfect interviewer or respondent performance; some researchers also like a separate code to differentiate respondent refusals to answer a question from questions unanswered for other reasons
   b. *Inapplicable information,* where the information does not apply to a particular respondent (e.g., length of hospitalization for those not hospitalized)
   c. "Don't know" answers may be treated as not ascertained or as a distinct category of missing data
2. Be consistent in assigning numbers; always use the same code for "not ascertained," "don't know," or "other" responses. The more consistent the code, the fewer the errors coders and programmers will make.
3. Make codes fit numbers in the real world when possible. Code numbers exactly (e.g., code a 45-year-old as 45). Also, number a list of responses in the order they appear in the instrument if there is no compelling reason to do otherwise.

When response alternatives are provided to respondents or the response form is highly structured, the code constructor's job is simply to assign numbers to the given set of answers and account for missing data. When respondents are asked to answer questions in their own words, however, the range of answers will not be fully predictable ahead of time. For such open-response questions, code development is an interactive process whereby the researcher identifies categories that emerge from the answers, as well as imposing order on the answers that are obtained.

The idea is to create categories that group answers that are analytically similar and to differentiate between answers that are different. If the categorization is too fine, the result will be many categories with only a few entries, which are hard to analyze and waste coder effort. On the other hand, large, gross categories may mask differences that are important. One criterion for a good code is that it must unambiguously assign each answer to one and only one code number. The other criterion is that it puts answers in analytically meaningful categories. How well the latter standard is met can be assessed only in the context of a clear plan for analysis.

In order to construct such a code:

- Have a clear idea about what characteristics of answers are of analytic significance. A good first step is to jot down the kinds of differences among answers to each question that are important from the researcher's point of view.
- Actually tabulate some of the answers from early interviews. Then construct a draft code for classifying those answers.
- Try the classification scheme on another 10 or 20 interviews; revise as needed.
- Have a separate code for "other" responses that do not fit the categories clearly and have coders make out notes recording these answers. The notes can be used to expand and clarify the code or add needed categories, as well as providing a record of answers included in the "other" category.
- The same kind of note should be used to allow coders to communicate problems or ambiguities in the coding rules to the researcher, who in turn should refine the definitions and policies.

These steps, together with an effective check-coding operation (discussed next), should produce an exhaustive and nonoverlapping categorization system that unambiguously puts each answer into one and only one place and that can be shared by coders, coding supervisors, and researchers who will analyze the data.

APPROACHES TO CODING AND DATA ENTRY

For many years, from the start of automated tabulation until the late 1970s or early 1980s, the path from respondent answer to data entry typically involved three steps: Answers were recorded in paper survey instruments; trained coders translated the answers into code numbers and wrote the numbers on special coding sheets; and keypunch operators punched the coded numbers onto 80-column IBM cards.

Even though these steps virtually never happen anymore, it is instructive to understand the purposes of the quality-control steps designed to minimize error with such procedures.

Interviewer Recording. There is no practical way to check whether or not interviewers record answers accurately (e.g., check the right box). It is good practice, though, to minimize the extent to which interviewers have to make coding decisions. Open-response answers are best recorded verbatim to be coded by trained, supervised coders. Interviewers should be instructed to write down all relevant information anytime a classification decision is at all unclear, so it can be reviewed and handled consistently at the point of coding.

Coding. Quality control of coding includes the following:

- Train coders, including having all coders code several of the same survey returns and then comparing results to make sure they are all coding the same way.
- Independently check-code a sample of each coder's work. This serves two purposes: It identifies coders who are making coding decision errors, and it identifies coding rules that are ambiguous and require clarification.
- A procedure should be established for coders to write notes about answers they are not certain they know how to code. These notes should be routinely reviewed by a supervisor. Such notes are an extension of the check-coding system, helping supervisors identify coders or codes in need of attention.

Data Entry. New computer technology has brought major changes in the data entry process. The old technology just described was geared to keypunch machines that recorded data by putting holes in IBM cards. Such machines were comparatively expensive and could do only one thing; hence only specialized keypunch facilities were likely to have them. Because of their relatively low cost and many potential uses, however, computers are omnipresent, and any such machine can be used for data entry. Hence, instead of sending data to a keypunching place, data can be entered by anyone almost anywhere. In addition, these machines can be programmed to improve the data entry process by

- permitting the entry of only legal codes in any particular field
- checking entries to make sure they are consistent with other previously entered data
- automatically ensuring that contingency questions are handled appropriately (i.e., when a series of questions is asked only of a subset of respondents, contingency instructions can be programmed so that fields for questions to be skipped will be filled automatically with the proper codes)

Although these checks do not identify data entry errors that do not violate the programmed rules, many data entry errors will be caught at a time when they can be corrected readily.

With this new technology there are two approaches to data entry that are now more common than the old three-step process. The first is a two-step process, often called direct data entry, wherein interviewers or respondents fill out paper survey instruments, and then coding and data entry are carried out in the same step. This coding and data entry can be checked by having a second person independently code and enter the data.

Another two-step option that deserves mention is optical scanning. The technology for scanning is evolving quickly. There are two ways that scanning is used to enter data.

Response alternatives can be bar coded, so that a person can enter numbers by passing a scanner over a bar next to the chosen response. The advantage of that approach is that a person without data entry skills can enter data.

Optical scanning of special sheets or forms, such as those used for standardized tests, has been available for years. That approach permits extremely low-cost data entry. The costs are in acquiring the equipment and, if special-purpose forms are needed, in setting up and printing the forms.

Historically, there have been downsides to optical scanning for use in surveys:

1. The forms were not user friendly, and survey researchers want questionnaires to be as easy to use as possible.
2. Creating special-purpose forms for relatively small surveys was fairly costly.
3. Significant missing data could result, particularly when unmotivated or unskilled respondents were asked to use them.

The last problem can be handled by doing hand checks of missing items to identify marks that the machines could not read. However, the promise for wider use of optical scanners in surveys probably lies in improved technology. Modern scanners are much more tolerant of imperfect marks than those in the past. They also can be used with a variety of formats, making them more adaptable to user-friendly survey instruments.

Scanning works well only with fixed-choice, precoded data, though progress is being made so written answers can be scanned for later coding by a coder. While the best equipment is still comparatively expensive, scanners are likely to play an increasing role in data entry in the future. Dillman and Miller (1998), Dillman (2000), and Blom and Lyberg (1998) provide good summaries of current scanning options and limitations.

The one-step processes are known as computer-assisted telephone interviewing (CATI), computer-assisted personal interviewing (CAPI), or computer-assisted self-interview (CASI), whereby interviewers or respondents enter answers directly into the computer and do the coding where needed. Paper and pencil are not used. Collecting data at an Internet site is essentially the same from a data entry perspective; answering the questions and data entry occur in the same step.

As discussed in Chapter 4, computer-assisted data collection is becoming increasingly common. For telephone interviews, each interviewer has a terminal at the telephone station. The question appears on the screen, the interviewer reads it, the respondent answers, and the interviewer enters the numerical value that corresponds to the answer into the terminal. That entry triggers a new question on the terminal screen. The computer can be programmed to accept only legal entries and to check the consistency of any entry with previously entered data, so the interviewer can clarify apparent inconsistencies in respondent answers. Laptops or other portable personal computers offer the same options for household interviews, and Web-based programs for collecting data perform in a similar way.

There are several attractions to all computer-based data collection systems:

1. The computer can follow complex question patterns that are difficult for interviewers or respondents in a paper-and-pencil version of a survey.
2. Information from previous questions or even previous interviews can be taken into account in question wording or the sequence of questions asked.
3. If inconsistent data are given, the interviewer can correct them immediately.
4. Data can be added to a data file ready for immediate analysis.

There are also some good reasons for not using such systems. Perhaps foremost among these is the lead time needed to program a computer-assisted protocol. The program must be error free if it is to be useful. Interviewers cannot deal with programming errors during an interview, as they can with typographical errors in a written schedule, and, of course, errors are even more problematic when respondents are entering their answers directly. Hence considerable time for testing and debugging must be allowed before starting to interview, though simple instruments with few skips will pose fewer problems for detecting program errors.

In addition, there is no quality control over data entry. There can be no checks on any data entry or on any coding decisions that interviewers make with a computer-assisted system, except to make sure that entries are legal codes and are internally consistent. Although keying error rates are relatively low, the greater concern is the quality of coding decisions (Dielman & Couper, 1995). Because of concerns about the lack of control over coding decisions, when open-response questions are asked, CATI and CAPI interviewers often record the answers verbatim into the computer for later coding. Nichols (1988), Baker and Lefes (1988), Saris (1991), Catlin and Ingram (1988), Nichols, Baker, and Martin (1997), and particularly the volume edited by Couper et al. (1998) provide good summaries of the characteristics, uses, and experience with computer-assisted systems.

DATA CLEANING

Once interviews have been coded and the data entered onto a tape or disk file, the data need to be checked. The most important check is to make sure the data file is complete and in order. In addition, every field should be checked to make sure that only legal codes occur. Even if there were checks built in at the time of data entry, it is good practice to make sure everything worked as planned by running a set of overall distributions. Of course, if checks were not done at the time of data entry, checks for internal consistency should be done as well.

When errors are found, the original source must be consulted and corrections made. (Note that this is not possible with a CATI, CAPI, or CASI system, because no hard copy is retained of the responses.) Because errors will be made during the correction process, checks should be run again. With large files, this kind of cleaning process is time-consuming and error prone. To the extent that errors can be caught at data entry, the reliance on postentry cleaning is reduced, which is highly desirable.

CODING AND DATA REDUCTION AS SOURCES OF ERRORS IN SURVEYS

Because coding and data reduction can take place in a highly supervised setting and can be checked thoroughly, there is the potential to have it be an almost error-free part of the survey process. Moreover, the costs of coding and data reduction usually should be a small fraction of the total survey cost.

When dealing with closed answers, the rate of error from data entry should be less than 1%. The level of error in the final data will be lower, of course, when those numbers are entered directly and 100% verified, so the transcription process itself is checked.

The reliability of coding open-opinion responses will vary with the quality of the question, the quality of the code, and the training and supervision of coders. If a researcher has a reasonably focused question, and if code categories are conceptually clear, one should expect coding to exceed 90% in reliability; that is, the coder and check-coder will disagree in the classification of fewer than 1 out of 10 answers. Coders who are not trained and check-coded appropriately create errors at considerably higher rates. Codes that depend on knowing complete definitions, such as occupational categories, health conditions, or specific crimes, may warrant special attention to coder training and check-coding.

The process of data entry can be nearly error free if it is verified. Although some individual operators are able to enter data at a remarkable level of accuracy, with error rates below 1 in 1,000 entries, one cannot routinely assume that data entry will occur at that level of accuracy.

The choice of the coding and data entry process will often be made for reasons other than the minimization of coding and data entry errors. The speed of file construction and the opportunity to catch errors during the interview are among the appeals of the CATI and CAPI systems, as are some of the strengths of involving a computer in specifying the wording and order of questions. Purely from the perspective of error reduction, however, the two-step process, whereby coders directly enter data and their work (coding and data entry) is 100% verified, may be optimal when a survey involves a significant number of coding decisions. No other system provides a true independent check on all coding decisions, as well as all data entry.

9

Ethical Issues in Survey Research

Like all social research, surveys should be carried out in ways designed to avoid risks to participants, respondents, and interviewers. This chapter summarizes procedures for ethically managing surveys.

As in all research that involves human subjects, the survey researcher needs to be attentive to the ethical manner in which the research is carried out. A basic guideline is that the researcher should make sure that no individual suffers any adverse consequences as a result of the survey. Moreover, to the extent that it is feasible, a good researcher also will be attentive to maximizing positive outcomes of the research process.

Almost all universities and most other organizations in the United States that conduct federally funded research have an Institutional Review Board (IRB) that is responsible for overseeing research involving human subjects. When research is proposed, the Principal Investigator must submit the proposed protocol for IRB review before beginning to collect data.

IRB review is designed to protect subjects, researchers, and institutions. In general, their greatest concerns are about research that involves some kind of risk to participants. "Research activities in which the only